AN ALGORITHMIC INTERPRETATION OF QUANTUM PROBABILITY

ALLAN F. RANDALL

A DISSERTATION SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN PHILOSOPHY
YORK UNIVERSITY
TORONTO, ONTARIO
JANUARY 2014

Abstract

The Everett (or relative-state, or many-worlds) interpretation of quantum mechanics has come under fire for inadequately dealing with the Born rule (the formula for calculating quantum probabilities). Numerous attempts have been made to derive this rule from the perspective of observers within the quantum wavefunction. These are not really analytic proofs, but are rather attempts to derive the Born rule as a synthetic *a priori* necessity, given the nature of human observers (a fact not fully appreciated even by all of those who have attempted such proofs).

I show why existing attempts are unsuccessful or only partly successful, and postulate that Solomonoff's algorithmic approach to the interpretation of probability theory could clarify the problems with these approaches. The Sleeping Beauty probability puzzle is used as a springboard from which to deduce an objectivist, yet synthetic *a priori* framework for quantum probabilities, that properly frames the role of self-location and self-selection (anthropic) principles in probability theory. I call this framework "algorithmic synthetic unity" (or ASU).

I offer no new formal proof of the Born rule, largely because I feel that existing proofs (particularly that of Gleason) are already adequate, and as close to being a formal proof as one should expect or want. Gleason's one unjustified assumption—known as noncontextuality—is, I will argue, completely benign when considered within the algorithmic framework that I propose.

I will also argue that, to the extent the Born rule *can* be derived within ASU, there is no reason to suppose that we could not also derive all the other fundamental postulates of quantum theory, as well. There is nothing special here about the Born rule, and I suggest that a completely successful Born rule proof might only be possible once all the other postulates become part of the derivation.

As a start towards this end, I show how we can already derive the essential content of the fundamental postulates of quantum mechanics, at least in outline, and especially if we allow some educated and well-motivated guesswork along the way. The result is some steps towards a coherent and consistent algorithmic interpretation of quantum mechanics.

To my past, present and future,

my parents, Gordon and Joan Randall, whose love and support have made me who I am,

my wife, Vee Ledson, whose love sustains me and who has become a part of me,

and my beloved son, Frank Ledson-Randall, who may one day be all that is left of me,

this piece of me is dedicated.

Acknowledgements

I would like first and foremost to thank my supervisor, Jagdish Hattiangadi, who has been with me and supported my efforts for the (too many) years it has taken me to complete this. He has provided feedback and encouragement and advice in so many ways. I would also like to thank the other members of my supervisory committee, Aephraim Steinberg and Dan McArthur, who have been supportive and helpful in numerous ways. In addition, I would like to thank the remaining members of my examination committee, Judy Pelham, Stanley Jeffers (for the feedback, but also for the touching keepsake passed on to me), and my external examiner, Jeffrey Bub, whose presence on the committee was truly an honour, and whose comments and feedback were extremely helpful.

I also must thank Martin Taylor, for many years now my scientific mentor. He supported me and helped me in a number of important ways, and while his contribution to the content was mostly indirect, his mark on this dissertation is incalculable, as his influence touches everything I do.

Thanks also to my wonderful wife, Vee Ledson, who tirelessly put up with "I'm almost finished" (for far, far too long) and helped immensely with various editing, illustrating and writing tasks (the introduction to Sleeping Beauty and much of the Chapter 1 introductory material are her words).

I also thank my parents, Gordon and Joan Randall, for being such wonderful grandparents to Frank and giving me the precious *time* I needed to finish the dissertation.

My brother Paul Randall had a direct and important influence on a number of the arguments contained herein, patiently critiquing my logic in a way that *only* he can do.

I thank Max Tegmark, for the exchange of ideas, but also for being a real leader and groundbreaker in the field of algorithmic ontology. Others have also inspired me with allied ideas, especially Jacques Mallah, whose influence appears in numerous places throughout this dissertation, as well as Robin Hanson, Travis Garrett, Peter Roosen-Runge, Max Rubin, Michael Weissman, and David Strayhorn.

Thanks to everyone else I have exchanged ideas with, who influenced this dissertation, including: John Sipe, Robert Hanna, Nick Bostrom, Robert Imlay, Robin Blume-Kohout, Gabriel Wendt, Hilary Carteret, James Hughes, George Dvorsky, Keith Henson, Tom Gee, Jim Brown, William Powers, Rick Marken, Henry Pietersma, Ian Hacking, Ann MacKenzie, Leslie Ballentine, Wojciech Zurek, David Wallace, Adrian Kent and David Albert. Thanks to Jeff Hunter and Hargurchet Bhabra for their early insights and help; had they lived, their mark on this dissertation would surely have been vast.

Finally, thanks to York University, its Faculty of Graduate Studies, and everyone in the Dept. of Philosophy, including Liz Bentham, Cristal Del Biondo, Melonie Ricketts, Rabia Sallie, Lorraine Code, Robert Myers and Claudine Verheggen for helping in so many ways to make this dissertation possible.

Contents

List of Figures

1 Introduction

1.1 Overview

There is a view of the universe in which all possibilities co-exist. Everything that *could* exist actually *does* exist somewhere in the great plenum of existence. This view of the universe is very old. When the possibilities referred to are fuzzy and ill-defined, it is a form of mysticism, which we will not be much concerned with here. But when the possibilities are clearly and distinctly specifiable, in a precise language, such a view forms the foundation-stone of philosophical rationalism, starting with Parmenides [149][175, Ch.15] in ancient times; and in modern times, with Descartes and the early modern rationalists [68, 67, 69, 207, 123].

There is another view of the universe, in which at least a great many (and perhaps all) possibilities co-exist. This view is based, not on *a priori* rationalism, but on the empirical science of quantum mechanics. It is the "many worlds" or "relative state" interpretation of quantum mechanics (or "MWI", for short), which posits the actual existence of a large number (perhaps an infinity) of other universes in order to explain some of the strange features of quantum mechanics. While quantum mechanics is an empirical theory, those who advocate the many-worlds interpretation tend to have at least some leanings towards rationalism. A hard-core empiricist, for instance, might well reject the idea of "other worlds" unless they were directly observable. Proponents of many-worlds, however, argue that we can infer the real existence of other worlds, if doing so explains the empirical phenomena in the simplest and most rational possible way.

Not everyone, however, agrees that the many-worlds interpretation explains the empirical data in the most rational possible way. Some argue that it has fatal flaws. I will attempt to use certain ideas from algorithmic information theory—developed from the viewpoint of a rationalist and computationalist epistemology—in order to address some outstanding philosophical problems with the many-worlds interpretation. I will especially focus on what I call the "Born rule objection", which claims that the MWI cannot adequately account for the "Born rule"—the actual formula used to calculate probabilities in quantum mechanics. This formula is not itself part of the mechanics or dynamics of the theory, nor is it directly derivable from it, but must be assumed as an additional

fundamental postulate.

Supporters of the many-worlds interpretation usually claim that the MWI's most convincing feature is that it derives from the dynamics of quantum theory alone, not requiring—like other interpretations do—any additional *a priori* metaphysical postulates. Yet, say the Born rule objectors, its inability to predict the Born rule shows that it must—just like any other interpretation—postulate something outside the raw mechanics of the theory. The objectors feel that if they are to be asked to believe in something that is as implausible (on the face of it) as multiple universes, then there had better be *no* such metaphysical loose ends at all.

I will argue, first of all, that the seriousness of this "loose end" has been greatly exaggerated. There are numerous proofs of the Born rule that derive it from the dynamics alone *plus* some assumptions. Sometimes these assumptions may seem quite dubious, but sometimes they may seem nearly self-evident (unfortunately, not everyone agrees on which is which!). Perhaps most significantly, Gleason derived the Born rule [91] by assuming a certain kind of "noncontextuality" of measurement: that when you do something to measure one thing, the probability rule will not be different from when you do the same physical thing, but in order to measure something *else* (or when you measure both things simultaneously). While this may be an assumption, it is one that a good many people would be willing to accept as a very intuitive and acceptable prior assumption, perhaps even self-evident.

Oddly enough, even though Gleason's proof is clearly analytically the strongest Born rule proof out there, most defenders of Everett against the Born rule objection do not consider Gleason to be relevant. Part of this is that he makes the unjustified assumption of noncontextuality. However, this would not seem enough to justify the wholesale dismissal of Gleason in the MWI probability literature, especially given that all the other attempts at MWI Born rule proofs make their own assumptions (some of them arguably far more dubious than noncontextuality). There is another reason, however, that Gleason is considered irrelevant. What the objectors are really seeking, it seems, is not a formal Born rule proof at all, but a derivation that shows why probabilities should obey the Born rule, *from the perspective of observers within the system*. From this point of view, even if we accept Gleason's noncontextuality assumption as self-evident, if the MWI (based on wavefunction dynamics *plus* some assumption about probability from the perspective of observers) predicts some *other* rule than the Born rule, then Gleason's proof becomes a *reductio ad absurdum* of the MWI. In other words, it proves the Born rule to be true, *counter* to the predictions of the many-worlds interpretation. This would, in fact, demonstrate the MWI to be inconsistent with the purely formal wavefunction. This is essentially, I believe, the position of many of the Born rule objectors.

Of course, this all assumes the objectors have a reasonable assumption to put forward about how

to compute probabilities from the perspective of observers within a formal system (which requires us to drop the fiction that the Born rule objection has anything to do with deriving the Born rule "from the dynamics alone"). The usual assumption made by the objectors is one of world-counting, or something closely related: the probability of an event is proportional to the number of worlds in which it happens. It is often simply overlooked that this is as much (or more) of an unjustified assumption as noncontextuality. However, if we could show that world-counting was no more *a priori* justifiable than Gleason's assumption—or even less so—then the whole Born rule objection, as it is generally framed, would fall apart (and to the extent that a problem with the Born rule was still an issue, Gleason's proof would be back on the table as a viable response for MWI advocates).

Aside from the Gleason proof, the Born rule can also be derived [79, 103] by first assuming that it is the "amplitude" of a quantum state that matters for calculating its probability as an outcome (an assumption that we will see later is actually closely related to noncontextuality). The amplitude of a wave is essentially the magnitude of the wave (so that, for example, the amplitude of a water wave on the beach would be the height of the wave). The objectors claim, however, that we cannot justify *assuming* that it is amplitude that matters (we happen to know that it *is* the amplitude that matters, but we only know this *a posteriori*—experimentally—not as an *a priori* feature of the raw dynamics of the theory, considered on its own terms). However, even when Born rule provers try to remove the assumption of amplitude dependence from their proofs, it has its way of sneaking back in and rendering the proof circular. However, this is not a slam-dunk for the objectors, either, because they are generally making their own assumption about what matters—namely, the world count.

It may seem, then, that we have the Everettians on one side, counting amplitudes, and the objectors on the other side counting worlds. However, while there is some sense to this dichotomy, it has not actually worked out like that, historically. Everett himself, as well as a great number of supporters of the MWI, have themselves *also* assumed world-counting. They have then been faced with the task of showing that counting amplitudes *amounts to the same thing* as counting worlds, usually in the limit of an infinite number of observations.

I will argue that—barring unrealistic expectations that we should have total knowledge of how the physical world generates our perceptions—any account of probability in quantum mechanics will necessarily have to make *some* kind of assumption about what fundamentally matters, and that amplitudes may even be a better starting point than worlds (worlds being much further removed from the basic mechanical elements of the theory). Hence, it could be argued that the whole motivation behind the Born rule objection is misguided.

On the other hand, I will also argue that the objection still has some clout, since, even though amplitudes are less of an "interpretational addition" to the dynamics than are worlds, this is no proof

that they are what matters. In addition, the objectors would be right to argue that if it *were* worlds that mattered, it would make sense to simply count them up, using the total number of worlds for our probability count. Yet, amplitudes do not "count up"—or rather it is hard to justify counting them up, since they can cancel each other out (an amplitude of +1 can cancel out an amplitude of -1, even though they both represent, on their own, the same possible outcome). Because of this possibility for "destructive interference", it could be argued that amplitudes have a serious problem as the thing that matters, *a priori*, for probabilities, even if we allow the assumption that they are clearly what matter most for the wavefunction dynamics.

I will develop this version of the Born rule objection as the version that is most worthy of a response (even though most of the actual objectors do not focus so much on this, being preoccupied with more easily refuted arguments that assume the priority of worlds). I will attempt, in Ch. 6, to sketch out a response to this version of the objection, by using algorithmic information theory (a version of information theory based on abstract computer programs and their program lengths). By starting with the idea that it is the information content of an outcome that matters for probabilities, rather than worlds or amplitudes, I will show that many of the (sometimes perplexing) features of quantum mechanics are to be expected, including destructive interference. This suggests that amplitudes may, in fact, be ultimately measures of algorithmic information content. I will argue that, if this were shown to be the case, it would give us a great deal more justification for preferring amplitudes over worlds, and a correspondingly greater confidence in the Born rule as the natural and expected probability rule for a many-worlds quantum ontology. In the process of developing these ideas, I will derive my own method—quite different from world-counting—for the *a priori* calculating of probabilities from the perspective of observers within a formal system (whether that system is quantum mechanics or something entirely different). I will show that, if we base this method on a rationalist and computational epistemology, a great deal of the fundamental postulates of quantum mechanics arise quite naturally, as more general rules for calculating probabilities based on complex-valued counts. I will even argue, continuing a process I started in [169, 172, 174], that perhaps the essential content of all the postulates can be derived in this way—at least in outline and if we allow a few reasonable educated guesses long the way.

1.2 Background

1.2.1 Quantum Mechanics

Quantum mechanics is one of the most successful physical theories in scientific history, in that it makes predictions that have been verified many times, to a high degree of accuracy. Without quan-

tum mechanics, and its incredibly accurate predictions, there would be no television, no microwave ovens, and no digital computers. While it is by no means perfect—most notably, it does not explain the behaviour of gravity—it provides an accurate foundation for most of the rest of physics.

In spite of its impressive experimental track record, however, there is something fishy at the core of this jewel of modern science. While in one sense, quantum mechanics yields precise experimental predictions, these predictions are in general statistical, predicting not necessarily what *will* happen in a given situation, but only the probabilities of different outcomes, for experiments that are repeated many times. Einstein was so uncomfortable with this indeterminism, that he was sure that quantum mechanics was somehow not the whole story, and that a better theory would one day restore determinism to physics. "God," he said, "does not play dice."

1.2.2 The Measurement Problem

The reasons for Einstein's discomfort went further than just indeterminism, however. For not only do the equations of quantum mechanics fail to predict which outcome will happen, they seem at times to be telling us that, in fact, all the possible outcomes are in some sense happening at the same time, in spite of the fact that when we look at the result of an experiment, we see only one outcome, not many. This conundrum is called the "Measurement Problem", and it pervades quantum mechanics.

It has been shown, time and again, in a wide variety of different laboratory experiments, that electrons and other subatomic particles, when emitted from some source—such as a light bulb or electron gun—do not move simply from one place to another, but actually seem to take all possible trajectories at once, somehow able to do multiple different things, and—in some sense that is not entirely clear—actually be at different places, at the same time. This is not to say that scientists actually can see an electron doing two different things at the same time, but rather that the equations of quantum mechanics that have been so successful at making predictions in the lab, seem to *describe* this kind of plethora of activity, whether or not it is "real". Yet, when the particle hits a measuring device, such as a Geiger counter or a photographic film, and we look to see where it landed, we do not see it in multiple places, we see that it landed at a single spot.

Indeed, quantum physicists can only infer that the particle was taking multiple paths at the same time because the statistics that emerge from repeating the experiment many times seem to demand it. I will not go into all the details here about why this is so—much of it will be discussed in later chapters—but physicists agree that there is *something* like this going on, although they disagree as to the exact nature of the process. The quantum equation that describes this "take all possible paths" behaviour is called the quantum "wavefunction", and when it describes multiple different

things happening at the same time, we call this a "superposition of states".

If we were to take the equation of the quantum wavefunction at face value, an electron in the lab is not only taking multiple trajectories at the same time—a superposition of states—but when it is measured, there likewise *ought* to be a superposition of measurement results, *if* the wave equation were truly telling the whole story. Since we never see such a superposition of results—we just see that the electron landed here or there, not somehow at both places at once—the question arises as to how the superposition described by the equations somehow turns into a single, determinate outcome.

It is tempting to say that the wavefunction somehow reduces down, or "collapses" when a measurement is taken. But this begs the question of *when* such a collapse is supposed to occur. Does it occur when the scientist becomes aware of the state of the measuring device? But this would give consciousness some kind of special role to play, and why would that be the case? On the other hand, the collapse might occur earlier, perhaps as soon as the electron impacts with the device. But then this somehow makes the measuring device special. Supposedly the device, whether it be a photographic film, Geiger counter, or some other contraption, is itself made up of subatomic particles that all obey the quantum equations, just like the electron that hits it. When two electrons collide, for instance, we get a new, more complex, wavefunction that describes all the things both electrons could be doing together, as they interact—we don't have one electron "collapsing" the other, or the two electrons collapsing each other. Why should a larger system not likewise continue to obey the quantum wavefunction, which demands a plethora of alternatives all co-occurring in superposition, not the single outcome that we actually see? The fact is, no matter how you slice it, if you are going to invoke a collapse of the wavefunction, you are stuck with the problem of where to place the collapse, since the quantum equations themselves have no such collapse in them.

This whole conflict between the superposition of the wavefunction and the much more singular experience we have when we look at a measuring device, constitutes the Measurement Problem. It might be easy to solve if we could somehow give "measurements" or "observations" some kind of special physical status, but there is no known reason to do this, and after almost ninety years of quantum theory, there is still no agreement as to how the terms "measurement" and "observation" should even be defined.

1.2.3 Interpretation

There have been many attempts to deal with the Measurement Problem, each producing a different "interpretation" of quantum mechanics. Some have posited changing the actual equations of quantum mechanics to avoid the problem. Strictly speaking, these are not really "interpretations" of quantum

mechanics, but are actually alternative theories, since they are capable, at least in principle, of making different predictions than quantum mechanics. However, if the changes made are minimal, and produce no predictions that differ enough from those of regular quantum mechanics to matter for practical, doable experiments, then it is traditional to study them as interpretations, rather than as alternative theories. There may even be some who would suggest that there are no "pure" interpretations of quantum mechanics at all, since the theory actually makes no sense on its own terms—it needs at least some minimal added machinery in order to make sense (I would not agree with this assessment, but it is one point of view that is out there).

My goal here is to address some of the outstanding issues within the many-worlds, or Everett, interpretation. While I will briefly discuss some of the other competing interpretations, this will be largely to give the reader a wider context for the discussion and to make my own biases and predilections clear; it is not meant to be an exhaustive survey of interpretations, and I do not make any claims to an adequate defence of Everett against his competitors. My primary goal is to address the claim that Everett's framework cannot properly deal with probabilities, even on its own terms. Therefore, the basic philosophical assumptions of the Everett interpretation will not be systematically questioned here.

One of the more metaphysically cautious interpretations is the "statistical interpretation" [11]. Here, the purpose of the wavefunction is not to directly describe reality, but rather to provide a statistical description of a reality that the wavefunction itself does not reveal to us. Under this view, the superposition is no more a description of multiple real things happening than is the statistical ensemble of classical statistical mechanics, where the movement of particles in a box is described in terms of an ensemble of possible boxes, yet no one believes there is a *real* ensemble of actual multiple boxes in the world. The ensemble is a statistical device. In this view of quantum theory, we are not necessarily barred from investigating the underlying reality—it is not a subject outside the realm of science—it is just not something that falls under the heading of "quantum mechanics", belonging more properly to the effort to find a successor, or deeper, theory.

Still others believe that it really is consciousness that somehow collapses the wavefunction, and that observers really do have a special role to play in nature. Wigner [233] defended such a view.

The most popular point of view, up until the 1970s at least, was Niels Bohr's "Copenhagen Interpretation", named after Bohr's home town. Bohr believed that we simply had to take measuring devices to be, for all practical purposes, non-quantum, or "classical" objects. Thus, a collapse occurs when a particle is measured by a *classical* (non-quantum) measuring device. Not that Bohr was really trying to create a dualistic theory of two kinds of matter, classical and quantum. His distinction was more pragmatic than that. In order to do laboratory experiments, he believed it was necessary to

make this distinction—one's experimental devices are the empirical starting point of our experiments, and simply must be taken as given. No scientific theory of how such classical objects behave was ever intended. The problem many (including myself) have with the perspective is that it, once again, side-steps the issue, this time by invoking a kind of entity that is outside the realm of scientific investigation.

The many-worlds interpretation posits that the plethora of seemingly contradictory things taking place in the wavefunction all *do* really take place, and that when we make a measurement, all the possible measurement results likewise also really take place, each one resulting in a different version of us in a different world or universe.

Some may ask, should we even care about which interpretation is "correct", if indeed any are? Indeed, if all we cared about were empirical accuracy, then there would indeed be no reason to worry about interpretation. This is actually, itself, another type of interpretation, and actually a fairly popular one. Here, the wavefunction is viewed as merely a "mathematical abstraction" that allows us to make certain predictions. It is not the role of science, it is said, to ask what the wavefunction means, or what it corresponds to in reality. For many, however (including myself), this is just to side-step the issue. I, for one, believe that science *is* about discovering how the real world works, and that interpretational issues are an important part of that process (and always have been). In any case, this is the approach I will be taking (although I will not have space here to defend it in detail).

Interpretation is important, not just to give us a comfortable way of thinking about a theory (although this too is important), but also because it can have profound effects on the path one takes to future theory formation. Quantum mechanics is clearly not the final theory of physics—at the very least, it is not compatible with Einstein's General Theory of Relativity, so it will very likely be replaced at some point by a new theory. How we put forward candidates for such new theories is highly influenced and guided by our interpretation of current theory. Indeed, it is even possible that we will never have enough information to definitively choose between competing interpretations until the theory itself has been superseded. Nonetheless, our development of a successor theory may well depend on our ability to first think clearly about the theory we have.

1.2.4 Schrödinger's Cat

A well-known example of the measurement problem is the Schrödinger's Cat gedanken experiment [194]. Although not a realistic example, it is well-known and easy to grasp, and its very unrealism forces one's intuition to grapple with the measurement problem.

Assume a sample of radioactive material is in a box completely sealed from the environment,

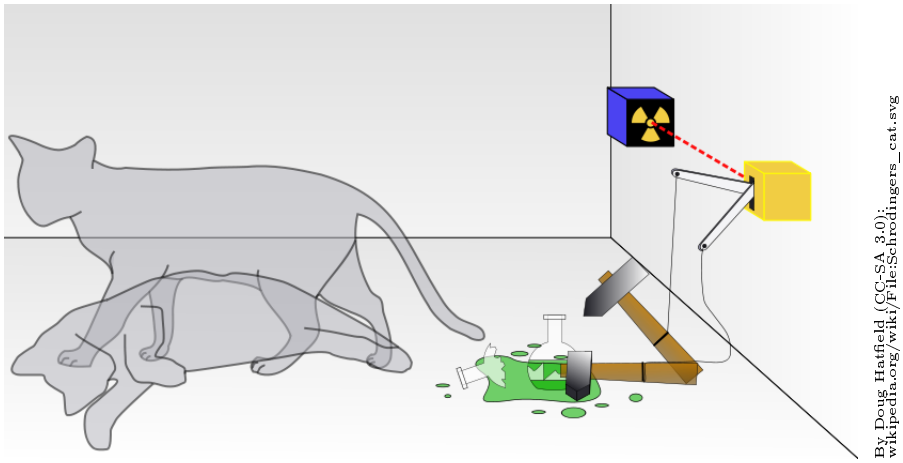


Figure 1.1: Schrödinger's Cat

along with a cat and a vial of hydrocyanic acid hooked up to the radioactive sample through a "diabolical device" (as Schrödinger called it) outfitted with a Geiger counter, and a hammer poised over the vial. The radioactive sample has a 50% chance of decaying in the next hour. The diabolical device monitors the sample for one hour, and if a decay is detected, it breaks the vial, killing the cat. Otherwise, the device does nothing and the cat lives.

Let's imagine (just to make things easier to talk about) that the radioactive sample is a single atom with a 50% chance of decaying within an hour, and that it is contained inside a box, which (like the larger box that contains the cat) has no interaction with the rest of the system. The diabolical device can detect the decay with 100% accuracy, but it does not look until the hour is up. (While these restrictions make the experiment less realistic, they serve to simplify the discussion, and do not change the essential character of the thought experiment.)

At the one hour mark, just before the device has had a chance to look, the radioactive sample is in a superposition of the decay state and the un-decayed state. After the device has acted, and if we do not invoke a collapse postulate, the wavefunction would seem to contain both living and dead cat, "mixed or smeared out in equal parts."

Here is the Measurement Problem in a nutshell: we need something like a collapse to occur, at the very latest, when we open the box and look inside, because we know from experience that we do not, in these kinds of situations, see cats in a superposition of being alive and dead. Yet, quantum mechanics gives us no reason to ever invoke a collapse up until that moment of observation. So how do we account for the collapse without making the act of observation somehow special? We could say the superposition collapses when the Geiger counter detects the decay, but then this surely makes the Geiger counter special, and quantum mechanics provides no reason to make measuring

devices special or different from anything else in the world.

According to Schrödinger, the problem lies in thinking that one can separate out subsystems from the larger system as a whole, and then suppose that entities in our model (in the wavefunction) describe real states of those subsystems. Schrödinger did *not* intend to make the claim that one could actually have a real world cat in a superposition of dead/alive states. Rather, his thought experiment was intended to highlight the problems associated with interpreting superpositions as representations of reality; most pointedly:

1. that subsystems in quantum mechanics, in general, have no state of their own, apart from the state of the entire system (the cat, then, has no state, blurry or otherwise),
2. that such subsystems in superposition, if taken as straightforward representations of reality, must necessarily include macroscopic superpositions (like the dead/alive cat),
3. that such a "blurred" picture of reality, since it describes a blurring of the entire system, and not of its parts, can be resolved merely by looking at the object, and therefore,
4. that the idea of superposition representing some kind of actual "smearing out" or "blurring" of actual things in the real world is untenable.

The thought experiment was intended, in other words, as a *reductio ad absurdum* of any naive realist interpretation of quantum mechanics. We naturally seek to build such realist models—in which we isolate parts of the model to represent parts of reality—because it is in our scientific tradition to do science this way. But quantum mechanics reveals a reality that is just not amenable to this method, according to Schrödinger: "Reality resists imitation through a model."

The problem with this conclusion is that, as Schrödinger himself admits, such a "blurring" of reality "in itself... would not embody anything unclear or contradictory." We call it "absurd" only because it is absurd to common sense; it is not logically or rationally absurd. And, as we shall see, not everyone agrees with Schrödinger that superpositions cannot be said to represent a "smeared out" reality. However, even if they do, Schrödinger is still correct that they do not represent a "smearing" of "parts" of the model. The "smearing", if there is any, is in the system as a whole, not isolated in its parts.

What if we *did* take seriously the idea that reality can accommodate macroscopic superpositions, or "smearing", of reality involving dead/alive cats? To do so, we would have to place a greater value on the scientific tradition of realist model-building than did Schrödinger—and likewise place *less* value on common-sense. At the end of the day, Schrödinger's Cat does not really demand that we reject either realism *or* common sense, but only that we cannot have both.

1.2.5 The Relative State Interpretation (Many Worlds)

In 1957, Hugh Everett III proposed a radical new solution to the measurement problem [79, 80], that took the idea of a reality that was "smeared out," on a macroscopic scale, seriously for the first time. He was primarily interested in applying quantum mechanics to cosmology—the study of the universe as a whole. The problem with the received wisdom at the time, dependent as it was on Copenhagenism, was that it was completely useless to cosmologists. While it might make practical sense in a laboratory experiment to divide objects into "classical" and "quantum", presumably the entire universe is made up of quantum objects, not classical ones. To apply a Copenhagen understanding of quantum mechanics to the universe as a whole would require an observer outside the universe to collapse it by making some measurement on it. But, of course, the universe being by definition all there is, there is and can be nothing outside of it to measure it—we can no longer invoke a classical measuring device to collapse the quantum wavefunction when the quantum wavefunction is everything that exists.

Everett wanted an interpretation that permitted sensible talk about the wavefunction of the entire universe, using only quantum mechanics, without invoking classical measuring devices. He wanted to preserve a realist approach to physics, and eliminate any special role for observation. Observers, he felt, should themselves just be quantum objects like any other, as should measuring devices. There should be no invocation of mental forces or mental states as fundamental entities—mental states should simply be emergent from the wavefunction dynamics. In other words, the fundamental postulate of Everett's interpretation is the idea of *wavefunction realism*: that the only thing represented in quantum mechanics, that actually exists in reality, is the wavefunction itself. Anything else is merely perceptual artifact.

A corollary to wavefunction realism is what is called the *no-collapse postulate*: the assumption that there is nothing like a real collapse of the wavefunction—for if the wavefunction is the only reality in the universe, we cannot allow the invocation of some external collapse mechanism. Any apparent collapse, therefore, is merely an illusion of perspective, like the apparent orbiting of the sun around the Earth.

The result of this thinking was the "relative state" interpretation of quantum mechanics. Imagine Schrödinger's cat in a superposition of dead/alive states. But recall that, by Schrödinger's analysis, we cannot talk like this. If we are to admit a blurry representation of reality—and Everett is insisting that we *do* permit the wavefunction to be a direct model of reality—then it must be the *entire system* that is blurry, not its individual parts. Schrödinger described this as a "disjunctive catalog" of conditional statements about reality, rather than a model of reality with blurry variables, as in:

"If the atom decays, then the vial breaks, and I observe a dead cat."

superposed with

"If the atom does not decay, then the vial does not break, and I observe a live cat."

So the disjunctive conditional catalog is now looking more like a catalog of possible histories for the entire universe. Since, from Everett's perspective, this "catalog" actually represents reality, we are forced to conclude that reality actually *is* a collection of such "histories"—or "worlds", or "branches" (Everett himself usually calls them branches).

So the wavefunction never collapses, and all the disjunctive histories/branches are equally considered to be parts of reality. While it is true that something like a collapse *seems* to occur—when we look at our measuring device, we see a particular outcome—Everett saw this, not as an actual collapse, but merely as a matter of perspective. Since, as Schrödinger stated, we are barred from ever referring to "states" of subsystems, such references must always be "relative to" some specified state for the rest of the system:

> There does not, in general, exist anything like a single state for one subsystem of a composite system. That is, subsystems do not possess states independent of the states of the remainder of the system, so that the subsystem states are generally correlated. One can *arbitrarily choose a state for one subsystem, and be led to the relative state for the other subsystem.* Thus we are faced with a fundamental relativity of states, which is implied by the formalism of composite systems. It is meaningless to ask the absolute state of a subsystem—one can only ask the state relative to a given state of the remainder of the system.[79, 80, emphasis mine]

So it is always the combined system that actually has a state. So, properly, we cannot say:

"The cat is in the dead state."

Instead, we must say:

"The cat is in the dead state *relative to* the broken state of the vial."

Of course, one could just as well state:

"The vial is in the broken state *relative to* the dead state of the cat."

The point is that the "broken" state for the vial subsystem is correlated with the "dead" state of the cat subsystem. The correlation goes both ways; there is nothing special about the cat as opposed to the vial. Note that we say that a particular *state* in the first subsystem is taken "relative to" a particular *state* in the second subsystem. The cat subsystem (of the overall cat+vial system)

can only be said to have the state "dead" *relative to* the "broken" state for the vial (not relative to the vial itself).

Of course, "cat+vial" is not really the "entire system"—and neither, ultimately, is "cat+vial+box", or even "cat+vial+box+experimenter". The "cat+vial" system only has the state "dead+broken" *relative to* some state or other for "the rest of the universe". So one could always make the argument that what we *really* should say is something like this:

"The cat+vial has the dead+broken state
relative to the such-and-such state of the rest of the universe."

However, given that we cannot always determine a precise state for "the rest of the universe", it is reasonable to allow more restrictive talk about states of smaller systems, so long as they are reasonably isolated (uncorrelated) with the rest of the universe, at least with respect to the states/variables that are of interest to us.

Of course, in the cat experiment, there *is* another subsystem of interest to us, and that is the experimenter himself. After the observation of the dead or alive cat—assuming the experimenter is himself just another quantum subsystem of the universe—then he himself will be in a superposition of seeing a dead cat and seeing a living cat. There are now, in a sense, two observers, each seeing something incompatible with what the other is seeing. No one observer ever sees a superposition of different outcomes, because the observer himself is in a superposition of seeing different outcomes—if this sounds contradictory, that's because it is, in a way. More precisely, we should say that it is an arbitrary matter of semantics whether we say that there is:

1. a *single observer* in superposition of having two independent experiences, *or*
2. *two observers* superposed in one system, each having an experience independent from the other's.

There is no answer to the question "is there one observer or two?", according to Everett:

> At this point we encounter a language difficulty. Whereas before the observation we had a single observer state afterwards there were a number of different states for the observer, all occurring in a superposition. Each of these separate states is a state for an observer, so that we can speak of the different observers described by the different states. On the other hand, the same physical system is involved, and from this viewpoint it is the same observer, which is in different states for different elements of the superposition (i.e., has had different experiences in the separate elements of the superposition). In this situation we shall use the singular when we wish to emphasize that a single physical system is involved, and the plural when we wish to emphasize the different experiences for the separate elements of the superposition. [80, p 68]

Some [7, 4, 15, 111] have tried to make sense of the idea that *only* the singular meaning is coherent. Their point seems to be that the "observer" is still a single "physical" system, albeit in superposition,

so there can only be a single observer. But, of course, it violates wavefunction realism to speak this way, giving some kind of absolute status to a subsystem, which is all an observer is in Everett's interpretation. Remember that subsystems do not even have states of their own, and are not even in superposition, on their own terms, but merely as part of the superposition of the entire system. Moreover, it is entirely arbitrary how one factors the wavefunction of the entire situation into "subsystems" in the first place (one could say the same thing for classical systems, for that matter). Hence, to insist that there is some kind of physical unity to the subsystem *on its own* that mandates the singular point of view does not seem consistent with Everett's program. There is no "absolute observer" here. The notion of an observer is emergent from the wavefunction, not a presumed entity within it. It is the wavefunction, in other words, that is absolute, not observers.

Still, we need a way to distinguish between these two different senses of "observer", in order to avoid confusion. Everett's suggestion that we use the singular "observer" and plural "observers" is not entirely satisfactory, as one can legitimately talk about multiple observers of type *(1),* or focus on a single one of the observers of type *(2).*

Definition 1.1. Hence, I will define:

1. the first (singular) kind of observer as the *observer-system,* or *physical observer*, and
2. the second (plural) kind of observer as the *individual observer*s, or *experiential observers.*

Since the observer-system cannot strictly be assigned a state, we need to speak of each individual observer as relative to an individual state for the rest of the system, usually corresponding to an experimental outcome. For instance, we can say that the experimenter has a particular experimenter-state only *relative to* a particular cat state:

"The cat is in the dead state *relative to* the experimenter seeing a dead cat."

Note that it is not really correct to say that the experimenter is in superposition of seeing a dead and living cat (although we may speak imprecisely like this when the context of the statement is clear). Strictly speaking, it is the above correlated subsystems, with their relative states, that are in superposition with each other:

"The cat is in the dead state *relative to* the experimenter seeing a dead cat."

superposed with

"The cat is in the living state *relative to* the experimenter seeing a living cat."

So the correlated relative states take the role of Schrödinger's "disjunctive conditionals". And the superposition of all such correlations takes the role of Schrödinger' disjunctive catalog—in other words, the entire wavefunction.

Note that Everett's interpretation is not giving privileged status to the observer or experimenter—quite the opposite; the observer is just another arbitrary subsystem within the *only* entity that has a non-relative state: the entire universe. Nonetheless, if we wish to account for our experience, we will need to choose states that are relative to whatever particular state we happen to find ourselves in. Our choice of such states may be ultimately arbitrary, but in practice we will need to make such choices, since what we are all really interested in is ultimately ourselves. Thus *before* we open the box and look inside to see if the cat is alive or dead, it makes sense to consider—from our personal perspective—that the cat is in a superposition of dead/alive states, since once we look, there will be *two* versions of us (one that sees a dead cat, and one that sees a living cat). However, once we look inside, since the two versions of us are experientially (but not physically) incompatible with each other (they do not share experiences or mental states), we *effectively* split in two (or more correctly, our entire universe splits in two, since that is ultimately what is in superposition, not mere subsystems). It is tempting to say that *subjectively,* there is a 50% chance we will see a dead cat when we open the box, but *objectively* both things happen, each in a different effective "world". However, for reasons that will become clear later, I consider use of the subjective-objective distinction here to be confusing, and prefer to say that there is a 50% chance from a given observer's viewpoint or *perspective,* but that both outcomes happen from an *observer-neutral perspective* (the "God's-eye" viewpoint).

However, since any *one* of the observers in superposition can never experience *all* of the worlds, once an observers looks and observes one and only one outcome (apparently at random), they are henceforth compelled to be interested only in those states relative to *that* outcome. Again, from the God's-eye viewpoint, this choice is arbitrary, but it is nonetheless necessary from the viewpoint of an observer, if we consider that statements we make about reality will be—in the overwhelming majority of cases—statements about whatever branch of reality we happen to find ourselves on. This is not a fact about nature, but a fact about ourselves—it is the same principle at work that determines that the vast majority of statements we make about economics will be about Earth economics.

Everett did not, in his published work, go so far as to say outright that the individual observers in superposition are living in different worlds or alternate universes, although he came very close to effectively saying so, and, in my opinion, it would be decidedly contrived to interpret what he said in any other way. He spoke of co-existing "branches", rather than worlds, but there is no effective

difference between his conception of a branch, and that of a world, so long as one retains the principle of wavefunction realism, and the idea that a world/branch is perspectival—relative to an individual observer—and not a metaphysical thing in itself. Everett himself assented a number of times to the explicit characterization of branches as perspectival worlds, stating that:

> Each individual branch *looks like* a perfectly respectable world where definite things have happened. [82, p 276, emphasis mine]

In any case, according to Peter Byrne, Everett's biographer [39, 40], Everett himself clearly intended his interpretation to be taken this way, but toned down the language under pressure from his supervisor, John Wheeler. Wheeler had shared earlier drafts with some of the Copenhagen circle (including Niels Bohr and Alexander Stern). The Copenhagenists had lashed out against it with enough force to put Wheeler into damage control mode. He attempted to convince them that Everett's interpretation was not, after all, as counter to Copenhagenism as it at first seemed, but was actually compatible with it, and intended as a generalization of it. Wheeler then convinced Everett to excise from his thesis much of the unorthodox language that had displeased the Copenhagenists [39]. Byrne reproduces a passage from an early draft of Everett's thesis that shows much more clearly that it is a many-worlds conception of relative states:

> The price, however, is the abandonment of the concept of the uniqueness of the observer, with its somewhat disconcerting philosophical implications.
>
> As an analogy one can imagine an intelligent amoeba with a good memory. As time progresses the amoeba is constantly splitting, each time the resulting amoebas having the same memories as the parent. Our amoeba hence does not have a life line, but a life tree. The question of identity or non identity of two amoebas at a later time must be rephrased. At any time we can consider two of them, and they will have common memories up to a point (common parent) after which they will diverge according to their separate lives after this point. It becomes simply a matter of terminology as to whether they should be thought of as the same amoeba or not, or whether the phrase "the amoeba" should be reserved for the whole ensemble.
>
> We can get a closer analogy if we were to take one of these intelligent amoebas, erase his past memories, and render him unconscious while he underwent fission, placing the two resulting amoebas in separate tanks, and repeating this process for all succeeding generations, so that none of the amoebas would be aware of their splitting. After awhile we would have a large number of individuals, sharing some memories with one another, differing in others, each of which is completely unaware of his "other selves" and under the impression that he is a unique individual. It would be difficult indeed to convince such an amoeba of the true situation short of confronting him with his "other selves".
>
> The same is true... [if] one accepts the hypothesis of the universal wave function. Each time an individual splits he is unaware of it, and any single individual is at all times unaware of his "other selves" with which he has no interaction from the time of splitting.
>
> We have indicated that it is possible to have a complete, causal theory of quantum mechanics, which simultaneously displays probabilistic aspects on a subjective level, and that this theory does not involve any new postulates but in fact results simply by taking seriously wave mechanics and assuming its general validity. The physical "reality" is

> assumed to be the wave function of the whole universe itself. By properly interpreting the internal correlations in this wave function it is possible to explain the appearance of the world to us (classical physics, etc.), as well as the apparent probabilistic aspects. [81]

I think, therefore, that it is misleading to suggest that Everett was putting forward something that was agnostic towards the real existence of other worlds. He did not see worlds, *per se*, as existent metaphysical entities onto themselves, but he did not see cats (or any subsystem, including himself) that way, either, and it would be contrived to suggest that he was agnostic towards the reality of cats. I will therefore use the term "relative state interpretation" and "many-worlds interpretation" interchangeably.

There are some variations on the MWI which have sought to retain the elegance of the no-collapse postulate, while jettisoning (or at least making optional) the multiplicity of real worlds (for instance, possibly the consistent/decoherent histories [100, 145, 88] and many-minds [7] approaches). In my opinion, these alternatives generally either fail to meaningfully distinguish themselves from the MWI in the first place, or lack the commitment to wavefunction realism that is the hallmark of Everett's entire approach.

The consistent histories approach, for instance, replaces "worlds" with "histories", but it is unclear—to me, at least—how this avoids being a variation on many worlds. Calling something a "history" instead of a "world" does not change anything of substance, any more than calling it a "branch". The "many minds" approach is more clearly distinguished from the MWI, but only gets around a really-existing multiplicity of worlds by positing a really-existing multiplicity of "minds"—whatever *they* are—thereby giving minds just the kind of exalted status in physics that Everett's wavefunction realism was from the beginning trying to avoid.

One of the most common objections to the MWI is not technical at all, but based merely on its common-sense implausibility. Proponents of the MWI generally feel that any apparent implausibility is a small price to pay for the elegant solution it provides to the Measurement Problem. Some detractors, however, find the idea of a universe constantly splitting into multiple copies of itself so implausible that they feel it can be rejected largely on that basis alone, regardless of its intellectual merits [157, 17, 230]. "Reality is not to be trifled with and sliced up in this way," objects Polkinghorne [157], who also declares many-worlders as members of the "'Gee-whizz' school of science popularisers, always out to stun the public with the weirdness of what they have to offer."

It is difficult, however, to see how rejecting an hypothesis based merely on its degree of common-sensical plausibility can be made to square with rational thought, the scientific method, or, indeed, with the actual historical record of successful scientific theories, many of which have been initially highly counter-intuitive. Such objections are, in my opinion, mere appeals to emotion, anti-scientific on the face of it, and I will therefore not give them any further consideration in this dissertation.

There are, however, critics of the MWI who do *not* reject it on account of mere counter-intuitiveness, who would be perfectly happy, it would seem, to accept a multiplicity of worlds if such a device could truly solve the Measurement Problem. Such critics disagree, however, with the standard Everettian claim that the MWI is the logical end-result of wavefunction realism. They believe, instead, that the MWI does, after all, add additional metaphysical constructs to quantum theory, above and beyond the wavefunction. Of all these objections, perhaps the most persistent one—and certainly the one taken most seriously by MWI proponents themselves—involves the relationship between the wavefunction and probability.

1.3 Frequentism and Bayesianism

One cannot really frame a response to the Born rule objection—or even coherently make the objection in the first place—without adopting an interpretation of probability theory. Unfortunately, as it turns out, the interpretation of probability theory is almost as controversial as the interpretation of quantum theory! The popular view is that the interpretation of probability is a choice between two options: *(1)* an objective view, called "frequentism", wherein probabilities are considered to be the long-run relative frequencies of events, in the limit of an infinite sequence of observations, or *(2)* a subjective view, called "Bayesianism", wherein probabilities are simply subjective degrees of belief, constrained by rationality and logic in how we reason from them. I will argue in Ch. 4 that this is a false and misleading dichotomy. However, we still need to understand it, as it is currently standard thinking and has had enormous impact on the quantum probability debate. For instance, I believe it is because of the prevalence of this false dichotomy—generally accepted by both MWI supporters *and* detractors—that Gleason's proof of the Born rule [91] is almost universally dismissed as irrelevant to the MWI probability debate, a dismissal which I will attempt to show is entirely unjustified.

1.3.1 The Frequentist Approach

Everett presented a proof of the Born rule, from MWI assumptions, in his original paper [79]. This proof falls into the general camp of frequentism, since it presumes that probability is the relative frequency of occurrence taken in the infinite limit. Frequentism is usually considered to be an objective view of probabilities, which seems consistent with Everett's realism.

Everett's proof can be divided into two stages:

1. Stage 1 shows that, *if* it is the amplitude of an outcome that matters, or "counts", towards the calculation of its probability, then we are forced to use an amplitude-squared rule if we are

to recover the expected mathematical properties of a probability measure. This does not, of course, prove that amplitude *is* indeed what counts.

2. Stage 2 of Everett's proof attempts to show that counting amplitudes amounts to the same thing as counting observers, in the limit of an infinite sequence of measurements on identically prepared systems. This is *not*—as is frequently claimed—an attempt to derive the Born rule from the wavefunction dynamics alone, since it assumes that branches (or *observers* or *worlds*) are the appropriate *a priori* thing-to-count. Branch-counting is commonly seen as a very natural assumption from an Everettian point of view: after all, since Everett claimed that all branches are "equally real", why should any of them be more probable than any other? Hence, if Everett really could show that the Born rule amounted to branch-counting in the limit, this might convince a lot of people (if not everyone) that the Born rule flows naturally out of Everettian assumptions.

I agree, however, with most of Everett's critics, that stage 2 of the proof fails on basic mathematical grounds, namely that it treats infinite limits incorrectly (and, in any case, there are no literally infinite measurement sequences in the real world, anyway).

On the other hand, I will *also* argue that the branch-counting assumption is not, in fact, a natural consequence of Everettian assumptions, and so even if Everett *had* successfully done what he thought he had done—reduce the Born rule to branch-counting—this would not in any sense have been an *a priori* proof of the Born rule, although it would still have been of great significance, since it would have unified amplitude-counting with its major, and perhaps only significant, competitor.

1.3.2 The Bayesian Approach

There is widespread dissatisfaction these days with frequentist conceptions of probability, and Bayesianism is usually considered to be its primary competitor for a foundational theory of probability, and there is some support for a Bayesian view of quantum probabilities [45, 3, 46, 43]. The Bayesian view treats probabilities as subjective, and not inherently about the real world at all, referring only to subjective degrees of certainty over one's purely subjective beliefs. The truth or falsity of such beliefs, in the Bayesian view, is not necessarily material to probabilities at all. Some Bayesians are stricter about this than others. Some may allow for the possibility of truly objective probabilities as one kind of probability within the Bayesian framework, while others may insist that all probabilities are necessarily subjective in nature.

Most Everettians have tended to be frequentists. However, one of the most popular recent attempts to derive the Born rule from Everettian assumptions falls broadly within the Bayesian camp, the proof of Deutsch and Wallace [226, 225, 71], which is based on principles of decision theory. Here, the Born rule is said to follow from the wavefunction dynamics, plus the MWI, *plus* some basic assumptions about rational belief and decision-making. It remains to be seen how many will be convinced by this proof, but see [84, 132, 14, 90, 164] for some responses.

In spite of its decision-theoretic framework, the basic nature of Wallace's proof is really not so different from Everett's stage 1. The key assumption made is that branch-counting is invalid (Wallace has verbal arguments, but no proof, for this assumption). This is not so different in spirit from Everett's stage 1 assumption that we must count amplitudes. Instead of assuming amplitude counting, Wallace is assuming that we can reject its main competitor. However, I will argue in §3.3.10 that Wallace not only fails to adequately discount branch-counting, but that his proof is no stronger or more convincing than Gleason's proof [91] (and could even be construed as a decision-theoretic version of Gleason). Hence, if Gleason's proof is irrelevant to the MWI Born rule debate (which seems to be the prevailing opinion) then Wallace's proof should not be considered any more relevant. However, in Ch. 8, I will disagree with the prevailing opinion, and argue that—under my own unproven, but I think well-motivated, assumptions in that chapter—both Gleason and Wallace can serve as Born rule proofs for the MWI, since the assumptions used in both proofs arise very naturally in an algorithmic context.

1.4 Branches versus Amplitudes

It seems, then, that the Born rule objection and the major responses to it—whether frequentist or Bayesian—are simply assuming opposing axioms, leaving little room for agreement at the current time. I would suggest, then, that any MWI Born rule proof that is based on amplitude-counting, in whatever form, will be open to attack by the either of the following counter-claims:

1. *Branch-counting (or world-counting, or observer-counting[1]) is better*: It is widely felt that *branches* ought to be what counts, *a priori,* not amplitudes. The term "branches" is a general term that can refer to worlds, outcomes or observers, since all of these can be said to "branch" in a many-worlds interpretation. The notion of counting observers in the relative-state interpretation was originated by Everett [79] in his original paper, while Graham [95] is perhaps the originator of the idea that world-counting is fundamental.
2. *Amplitudes are not the kinds of things that add up to probabilit*ies: Unlike with, say, sound or water wave amplitudes, it is unclear exactly *what* quantum amplitudes are supposed to represent, and (as we will see later) they have some decidedly strange properties that make their use for "counting" in probability calculations quite odd—namely, their ability to interfere with each other.

It is not entirely clear what the justification for #1 really is supposed to be. Born rule objectors frequently just assume that world-counting is, from an *a priori* perspective, self-evident. The Born

[1] I have lumped observer-counting and world-counting together here as "branch-counting", and will continue to do so where appropriate; however, as we will see in later chapters, they are not exactly the same thing, and will have to be treated separately in some contexts. It is less common to talk about "outcome-counting", although certainly both world-counting and observer-counting are really, at the end of the day, based on outcome-counting (since it is, after all, only outcomes that we *actually* empirically count in laboratory experiments).

rule, on the other hand, they seem to view as illegitimate because it counts branches unequally, even though they are supposed to be "equally real". The problem with this view is that, since worlds or observers are not primitive elements of the wavefunction, but appear to be more like emergent properties, they may all be "equally real", but this does not mean that they are the "real things" of objective reality (the "ontic entities" of our ontology). The claim to "equal" reality for various branches just means none is more real than any other. The idea that we can leap from this to the postulate that branches are the actual "real things" of the world, that we can count up, is dubious and has, in fact, hardly been defended at all. This is especially true given that amplitudes arguably *are* the "real things" in any ontology based on wavefunction realism. Wavefunctions are, after all, waves of amplitude, or at least of whatever it is amplitudes represent. "Branches" do not even appear in an analytic description of the wavefunction.

The evolution of the wavefunction is said to be "unitary". I will define more precisely what that means in Ch. 2, but for now, we will just note that this unitarity is essentially the conservation of amplitude structure. And since amplitude-counting yields the Born rule, we are quite justified in claiming that the Born rule can be derived simply by assuming that we need to count whatever it is that is the *conserved quantity* of the wavefunction: in other words, whatever it is that the wavefunction "carries" in the way that classical waves carry energy, which is *their* conserved quantity. For a classical wave, we "count" amplitudes (by integration) in order to compute how much "real conserved stuff" (energy) it represents.[2] Probabilities also have to be "conserved", as well (total probability will alway be unity, or 100%) so there is a natural tendency to look for a conserved quantity when looking for any countables for probability calculations.

Thus, amplitudes really *are* a very natural thing to count (even aside from Gleason's proof). So the objectors really need to do far more than simply present an alternative to amplitude-counting—especially given the derivative and emergent nature of their actual proposed alternative. This is the reason that objection #2 is actually quite important to making the Born rule objection viable. It argues that, while amplitudes may have many mathematical features that make them an obvious candidate for counting, they are simply not the kinds of things that make sense to count up, for the purposes of computing probabilities. This is largely due to the existence of destructive interference in wavefunction dynamics. In classical probability theory, countables cannot "interfere" with each other in this way. If there is *at least* one way for an event to happen, it will *with certainty* have a non-zero probability—its chances of happening cannot be cancelled out by the existence of still other ways that the same event could also happen. The argument is that this just does not make

[2]The distinction between discrete counting and continuous integration is not important here, so long as the reader keeps in mind that I am including integration as a kind of counting or summing up. For quantum waves, amplitude counting can take the form of discrete counts or continuous integration.

sense, from the perspective of probability theory.

Of course, there is still no point in arguing against amplitude counting if one cannot even suggest a *possible* alternative. Hence, objection #1 is also needed here: we need an alternative, observer-oriented way to count. Generally speaking, it seems to be the case that any such alternative to amplitude counting asks us to accept a probability measure that is in some way *subjective*. The problem is—as we will explore in the next section—that there are *at least* two entirely different ways that a count-based probability measure can be called "subjective", and much confusion emanates from the inability to distinguish between these different senses of "subjective".

One of the most important points I will be making, which underpins almost everything else in this dissertation, is that a simple counting of branches (whether conceived of as "worlds, "outcomes" or "observers") is inappropriately subjective. Wallace gives his reasons for considering branch counting to be invalid:

> 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'... 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. [227, p 255]

I believe that a misunderstanding of this point is the primary source of confusion about the MWI. In fact, while I agree with Wallace's basic point above, I will argue (in §3.3.10) that even Wallace ultimately gives the wrong argument against branch-counting. When he says that "there is no such thing as a 'branch count"', I would argue that he should really be saying that "there is no such thing as an *analytic* branch count." We will see later, however, that there *is* a such thing (counter to Wallace) as a non-analytic (*i.e.,* synthetic) yet still *objective* branch count. Each "version of me"—*i.e.* distinct consciousness—that exists after an observation is simply counted as one branch. However, these branches are perspectival artifacts of the observer's perception, not objectively existing (ontic) entities. Hence, they cannot serve as countables for an objective probability measure, even though they are objectively discernible. The problem for Wallace here is that he is not adopting an objectivist view of quantum probabilities, and so he is unable to insist that countables must be objectively existing entities. For that, one would require a commitment to a metaphysical foundation for probabilities, which is contrary to Wallace's whole decision-theoretic approach. Thus, he tries to argue that branches are not even definable at all. But his appeal to the "physical" vacuity of the branch count does not really help him to dismiss the branch count as a basis for probability—since he is adopting a (potentially) subjectivist approach to probability, the branch count surely does not *a priori* have to be "physical" in order to matter. Wallace would have an argument here only if

he could show that the branch count was not only non-physical, but could not even be objectively calculated—not even from the perspective of a particular observer. To do this, he would have to show that, given a complete description of the observer and environment, there is no way (even in principle) of objectively determining how many distinct post-measurement consciousnesses continue that observer's pre-measurement conscious state. Such a demonstration would seem to me to be highly implausible, and in any case, Wallace makes no attempt to argue for anything like this.

I will be dismissing branch-counting, nonetheless, for nearly the same reason as Wallace: I agree that there is no single analytic "branch count". However, I will argue that one must adopt a thoroughly metaphysical take on quantum probabilities as *objective chances*, in order for this observation to be meaningful, and properly grounded. For this, a purely decision-theoretic probability theory does not seem to be the proper tool.

The idea of branch-counting seems very intuitive to most Born rule objectors (and even many Everettians). Indeed, Everett himself supported it in the form of observer-counting. It might seem, for example, that if $1/4$ of the post-measurement observers experience outcome $\mathcal{X}$, then the pre-measurement probability of $\mathcal{X}$ should, in a many-worlds ontology, be $1/4$. In spite of Everett's support for this reasoning—in stage 2 of his Born rule proof—this kind of scheme does not actually follow from his own fundamental assumptions, and is, in fact, in gross violation of them, since Everett's most fundamental postulate is that of "wavefunction realism" (not "world realism" or "observer realism"). Worlds, in fact, in the relative-state interpretation, are properly considered as perspectival artifacts, relative to an observer. In other words, they are *synthetic* in nature (definable only in terms of experience), unlike (for example) *amplitudes*, which have a precise *analytic* definition, even if we were to completely ignore the idea that there might be experiential observers hiding in, or emergent from, the wavefunction.

1.5 Objectivity and subjectivity

Sometimes, it will be more accurate to refer to the entire class of all synthetic counting methods as "observer-dependent counting". This class will include some methods that do *not*, in general, produce the same numbers as branch-counting. For instance, we might consider adding the following alternatives to our list of potential observer-dependent countables, alongside the counting of branches, worlds, outcomes or observers.

3. *observer caring*: Greaves [97] suggests the idea of tying the weight we give to a particular branch to the amount we "care about" that branch.
4. *observer fatness*: Albert [6] goes one better with such "caring" measures by suggesting a variant: one in which the weight of a branch depends on the mass of the observer in that branch, since

you should "care more" about worlds in which you are fatter (since there is more of you in those worlds).

Both of these suggestions seem inherently implausible—as they both imply purely *subjective* quantum probabilities, and quantum probabilities do not seem to be things that depend on what we care about or believe. They seem to be objective in nature. If some outcome has a quantum probability of 25%, then this number will surely be the same, even for an observer who has no capacity to understand the idea of probability, care about the outcomes, or even conceptualize the categories involved.

But aren't branches, being synthetic and observer-dependent, just as subjective as caring or fatness? But I think it fair to say that most people would find it inherently much more plausible to say that probability depends on the world count than to suggest that it depends on how much I *care* about certain worlds, or my fatness. (The fatness measure, of course, was not meant to be serious, but rather was meant to show what is wrong with the whole idea of caring measures.) What we need to ask now, is whether the problem with such measures is their *subjectivity*, and if so, whether or not this also eliminates world-counting and observer-counting—or whether there is something wrong with the whole subjective-objective distinction, as we have been drawing it.

I will argue that, while it is possible to regard branch-based measures as "subjective", if so, they are subjective in a completely different sense than caring and fatness. But first, let's establish what we mean by the subjective-objective distinction, in the first place. Different people mean different things by this distinction, and the failure to properly define the distinction can lead to great confusion.

I will define three different levels of objectivity/subjectivity for probabilities:

1. "Strongly objective", or "metaphysical" probabilities are not dependent for their value on the state of knowledge or beliefs of the observer.
2. "Weakly objective/subjective", or "epistemic" probabilities are dependent for their value on certain pre-assigned "priors" (prior probabilities) that *correctly* quantify the observer's relevant knowledge. All other probability calculations and inferences are made as if the probabilities were objective.
3. "Strongly subjective" or "psychological" probabilities are dependent for their value on certain arbitrarily assigned "priors" (prior probabilities), that may or may not rely on the observer's state of knowledge or beliefs. All other probability calculations are made as if the probabilities were epistemic.

Note that these are not general definitions of objectivity and subjectivity, but simply define how I intend to use these words, in this dissertation, as applied to probabilities.

There is a further distinction between two different senses of "subjective" that I would like to make with respect to count-based probability measures. A count-based measure can be subjective

in either one (or both) of two possible ways. Given that a counting probability will always ask us to divide the number of cases in a certain category (like the number of *red* marbles) by the total number of cases (like the number of *marbles*), I will classify these two kinds of subjectivity according to whether they subjectify the *numerator* or the *denominator* of the classical probability ratio.

Definition 1.2. "Denominator subjectivity", or "count subjectivity*":* the measure counts non-ontic, perceptual (or otherwise subjective) entities that are dependent on the observer's knowledge or beliefs; contrasted with "denominator/count objectivity", in which objective (ontic) entities, with an observer-independent existence, are counted.

Definition 1.3. "Numerator subjectivity", or "category subjectivity": the measure *categorizes* what it counts into categories that are an arbitrary choice of the observer; contrasted with "numerator/-category objectivity", in which the countables are partitioned into categories that the observer has no control over (meaning that there are reasons why those categories *must* be chosen, independent of the observer's preference for them, or even knowledge of them).

It is important to divide subjective and objective counting measures like this, as much confusion can arise from a lack of recognition of this distinction.

We may "care more" about one branch over the others all that we like, but this is not going to stop the branch with the higher quantum probability from being actualized for us, instead of the one we "cared" more about. This kind of all-out subjectivity would leave no room for the clearly objective nature of the vast storehouse of empirical results that have been collected over the years.

The "subjectivity" in branch measures is of an entirely different kind. Since branches are perspectival, they cannot be ontic entities, and we cannot thereby have count-objectivity. However, we could still *categorize* our counts according to branches, and we could have category-objectivity. This would require justifying the use of this category as observer-independent, and I do not believe that such a justification exists, but we certainly cannot rule it out *a priori*.

Note that we would become hopelessly confused if we got these two senses of "subjective" mixed up. This confusion is actually terribly common-place. In the context of branch-counting, it is extremely tempting to conflate "branch categorizing" (in the numerator) with "branch-counting" (in the denominator).

1.5.1 An Example: the four worlds

Let's assume for the sake of argument that we have decided to count "worlds", but in the sense of rocky planets, not Everettian worlds. We can certainly imagine perfectly ordinary (non-quantum) situations in which we would naturally count "worlds", and no one would consider it unacceptably

"subjective" to do so. Imagine you and I each have our own spaceships, and we play a game of Solar Hide-'n-Go-Seek. I tell you that I will hide on one of the Solar system's worlds (meaning rocky planets), but I won't tell you which one. You have no reason to think I am any more likely to choose any of these four worlds over the others, but, since you have to start somewhere, you decide to search Venus first.

Question. *As you sit in orbit over the planet Venus, what is the probability, from* your *point of view, that I am hiding on Venus?*

Most of us would accept that the probability is 1/4. This is clearly in *some* sense objective. No amount of "caring" about Venus, or "hoping" that I am hiding there, will make it more likely that I am actually there. On the other hand, you don't really know how I decided which planet to hide on. So the 1/4 probability is based partially on your ignorance of what I did. Thus, if you could somehow repeat the experiment many times, you would *not* at all necessarily find that I am on Venus one of every four trials. In fact, if you live in a deterministic universe, you might find that I hide on Mars 100% of the time—perhaps because that is my favorite planet, a fact about me that you did not know.

So the 1/4 result cannot be *completely* objective—in the sense of being entirely independent of your knowledge of the situation. Clearly, it is only "objective" if we allow it as given that I have no reason to prefer any world over any other. But all this really says is that *you* have assigned a "prior probability"—based on your own ignorance—of 1/4 for each world. Any inferences you then make from this *subjective* prior are *in other respects* objective, so long as they follow the rules of logic and the mathematics of probabilities. If your ignorance is *correctly* quantified by the number 1/4, then it is an epistemic, or weakly subjective probability. If not, it is strongly subjective, or psychological. While epistemic probabilities may still be argued to be fundamentally subjective, they clearly have more claim to objectivity than those based gratuitously on caring, hope or intuition.

Assuming that you really do have no knowledge that would allow you to prefer one world over another, your conclusions are all *objective inferences* from your current knowledge. But your knowledge is not itself objective, since it depends on you. On the other hand, it is possible for you to *believe* that there is an epistemic probability of 1/4 that I am on Venus, and you might be mistaken. Someone else could come along and *remind* you that Mars is my favorite planet, and then you might realize that the probabilities are not all equal, after all. (This leaves you with the sticky problem of *how* to correctly quantify your ignorance, which may not be doable with precision, but we can accept this ambiguity as unmysterious.)

On the other hand, if you simply decided that you wanted to assign a prior of 7/8 for Venus, simply because Venus was *your* favorite planet, your resulting conclusions about probabilities, even

if they be impeccably and objectively inferred from your priors, can *not* be called epistemic. If your priors are based on hopes, desires, likes and dislikes, rather than being quantifications of your level of *knowledge*, then your probabilities are fundamentally more subjective than epistemic probabilities. Later, we will call most of these kinds of probabilities *doxastic,* since such non-epistemological subjective priors are usually—if they are to behave like probabilities at all—based on prior *belief*, if not knowledge. However, the "strongly subjective" label will also apply to any inferences drawn from non-epistemologically justified priors.

One problem in interpreting quantum probabilities is that strongly objective (type 1) probabilities are rarely used in classical physics. Indeed, it would seem that, *in reference to single cases, probabilities other than 0% and 100% can never be strongly objective in a deterministic universe.* For if the universe were truly deterministic, and you could somehow repeat a particular (single-case) experiment, under *exactly* the same conditions (*i.e.*, with the entire universe in exactly the same state), then you would simply get exactly the same result every time. If you were to appeal to a repetition of the experiment that had the universe in even the teensiest bit different state, then your resulting probability might be other than 0% or 100%, but it would not be *single-case*, and it seems to be single-case probabilities that are at the root of most of the problems in the foundations of probability. In addition, quantum probabilities are not restricted to 0% and 100%, and they are (or at least appear to be) single-case probabilities.

Thus, our choice of world-counting in the Hide-n-Go-Seek example might be the objectively appropriate choice—*if* it properly quantifies our ignorance—but *if* the universe is deterministic, no single-case probabilities other than "0%" or "100%" can *ever* be strongly objective, so the Hide-and-go-Seek probabilities would seem to be clearly subjective to some extent.

Here is the problem: it is because quantum outcomes *do* seem to be genuinely nondeterministic, even when applied to single cases, that we get the possibility that they are strongly objective. In fact, if we believe in wavefunction realism, it becomes hard to deny that quantum probabilities *are* strongly objective. When one takes the many worlds viewpoint, however, a new subtlety emerges, which is that *all* the outcomes described in the wavefunction are given equal reality. There is, in fact, no absolute division of these results into "worlds". The very idea of a world is relative to a particular individual observer, and hence would seem to be subjective.

Hence, we have an undeniable tension here. In a sense, these probabilities must be strongly *objective*, since they remain the same, regardless of the observer's knowledge or belief (even if we were rabbits, the quantum probabilities would be the same, without our even having any conception of probability at all). This seems the height of observer-independence and strong objectivity. And yet, unless they are relative to an individual observer, such probabilities disappear into nothing,

making them (seemingly) completely subjective.

This tension must be resolved *first* if we are to talk sensibly about whether quantum probabilities are "objective" or "subjective". To help resolve it, we return to our planetary Hide-n-Go-Seek example. Assume for now that you *can* correctly quantify your knowledge, and so your probabilities are epistemic, and therefore weakly subjective. Assume further that we ask not the probability of my hiding on Venus, but the probability of my hiding on any planet that is less than $12,500$ km in diameter. This is a category. So, given that our ontic entities are worlds (planets), we decide to count worlds, grouping them according to this category. Given that our category is "size", this means that we need to take the count of worlds under the size limit, and divide by the total number of worlds. Since only the Earth is actually above the limit, the count in the numerator will be 3, and the probability will be $3/4$.

Again, this assumes that all worlds that meet this category requirement are equiprobable, and that worlds are the entities we are picking from (the things to count). If this is a correct quantification of our ignorance, then this probability is weakly subjective. However, if we accept that worlds really are objective countables, then this probability is only subjective in the numerator (category-subjective), not in the count.

Now what if you knew with certainty that I had based my choice of world on a sequence of two fair coin tosses? If we accept a coin toss as an ontic countable (that *heads* is objectively just as likely as *tails*) then our probability is still $3/4$, but is now strongly objective. At least it is objective in the count; but is it also objective in category? This can be a tricky question, because it depends on more than just the category itself. It depends on the question I am trying to answer. All categories might be called subjective *in some sense*, since there is no absolute compulsion to categorize in the first place. However, if I ask you the probability of drawing a red marble from a bag, my assumption of the category "red" is included in the question posed, and so does not render the probability subjective (although we will still say it is subjective *in category*). The resulting probability, however, is strongly objective. Of course, one might argue that coin flips are not ontic entities, and that taking them to be such is only a convenience to avoid having to fill in our missing knowledge. However, if you like, substitute an appropriate quantum event for the coin flip. It then becomes extremely difficult to avoid the probability's being objective (although, again, it remains subjective in category).

One might think, then, that unless we have a perfectly flat distribution, where all events are equally likely, that the numerator *must* contain reference to some category or other, and therefore all probabilities are subjective in category, even when they are strongly objective. However, this is not the case. Imagine someone asks you for the probability that the world I am hiding on is beautiful, or within 1000 km of you. These categories will be defined differently for different people, and so

there is a subjective element to them that does not exist for planet diameter. However, even then, these two examples are very *differently* subjective. Beauty is subjective in a way that is—in practice if not in principle—impossible to objectify. Distance to an observer, however, is only subjective because the location of the observer is not specified. This kind of subjectivity can be objectified by further analysis of the state of the observer. And *perhaps* even beauty can be so analyzed, if we only had enough information. I will not try to settle *that* question right now—our purpose here is merely to clarify the distinction between the possible senses of "objective" and "subjective", not to settle all resulting controversies (although some of the more important ones will be discussed in Ch. 4).

The failure to distinguish between all these very different senses of "subjective" is at the root of much confusion over the role of probabilities in Everettian quantum mechanics. It arises, I believe, out of the attempt to make a single dichotomy, "objective versus subjective", do duty for too many different ideas, and it has generated the further false dichotomy of "frequentism versus Bayesianism". Fortunately, there is already an additional distinction that clarifies a great deal of the confusion over subjectivity, that already has a long philosophical history, and, in spite of widespread rejection in the twentieth century, it is extremely well suited to quantum probabilities.

1.6 Analytic versus synthetic

So should we take quantum probabilities to be objective or subjective? While we could certainly *define* quantum probabilities as strongly subjective, based on their strong observer-dependence, this would be a very different sense of "subjective" than that defined in the previous section. I believe we need to very strongly distinguish between "subjective" and "observer-dependent". A statement or judgment that is observer-dependent is better called "synthetic", rather than "subjective", since it is possible for it to be actually quite strongly *objective* in the sense we defined above. Synthetic statements and judgments are contrasted, not with objectivity, but with analyticity. As a preliminary to defining these terms, however, it will also be useful to explain a few other related, standard epistemological terms.

Definition 1.4. We will define "knowledge" in the standard way, as "justified true belief". Thus, we will not require knowledge to be *certain*, so long as it is *rationally justified*. But justification alone is not enough, for if a justified belief is nonetheless false, it is not knowledge. Neither is it knowledge if it is true—and we fully believe it to be true—but our reasons for believing are misguided. While it is possible to debate the validity of this definition, it will do fine for our purposes.

Definition 1.5. A "necessary statement" is a statement that *must* be true—it cannot, even in

principle, be coherently imagined to be false. An "impossible statement" is the denial of a necessary statement. A "contingent statement" is one that could conceivably be true or false—it is possible, in principle, to imagine it either way.

Definition 1.6. All statements are either necessary, contingent or impossible. A "possible statement" is one that is not impossible (so it is either necessary or contingent). An "unnecessary statement" is one that is not necessary (so it is either contingent or impossible). A "noncontingent statement" is one that is not contingent (so it is either necessary or impossible).

Definition 1.7. "Analytic statements" or "logical statements" are logically true (or false), so that their truth values do not depend in any way on experience, being "observer-independent"—they are thus noncontingent (either necessary or impossible), since it is not possible, even in principle, to imagine their being false (or true, as the case may be). An "analytic truth" is an analytic statement that is true. An "analytic falsehood" is an analytic statement that is false. An "analytic judgment" occurs when an observer decides, by using logic alone—without appeal or reliance on experience in any way—that a statement is true or false.

Definition 1.8. "Synthetic statements" or "experiential statements" are statements that are dependent on experience; they are thus "observer-dependent" and *not* true (or false) by logic alone—they are therefore contingent, since it is logically possible to imagine, at least in principle, their being false (or true, as the case may be). A "synthetic truth" is a synthetic statement that is true. A "synthetic falsehood" is a synthetic statement that is false. A "synthetic judgment" occurs when an observer decides, by nonlogical means—appealing to or relying on experience in some way—that a statement is true or false.

In this dissertation, "synthetic" is the same as "experiential", and "analytic" will essentially mean "logical", which will not be taken as particularly distinct from "mathematical" or "computational". So any true mathematical statement is no less analytic than a logical tautology. Any computer program that we can construct can also be said to be an analytic truth, even though we normally express programs constructively, and not necessarily as statements (but just attach the phrase, "the following program exists..." to the front of any program, to make it a statement).

Definition 1.9. An "*a posteriori* judgment" occurs when an observer decides, in an observation-dependent way—*i.e.,* by way of particular experiences—that a statement is true (or false), whether or not the judgment is made with certainty. An "*a posteriori* statement" or "empirical statement" is one whose truth can most certainly be known through "observation-dependent" means, since it follows from *particular* experiences, or "observations". An "*a posteriori* truth" is an *a posteriori* statement that is true. An "*a posteriori* falsehood" is an *a posteriori* statement that is false.

Definition 1.10. An "*a priori* judgment" occurs when an observer decides, in an observation-independent way—*i.e.*, without appeal to particular experiences—that a statement is true (or false). An "*a priori* statement" is one whose truth can most certainly be known through "observation-independent" means, since it does not follow from any particular experiences, or observations. An "*a priori* truth" is an *a priori* statement that is true. An "*a priori* falsehood" is an *a priori* statement that is false.

Definition 1.11. From the definitions already given, a "synthetic *a posteriori*" or "*a posteriori experiential"* statement is one that is "experiential, and knowable only by way of *particular* experience"—*i.e.* observa*TION*-dependent (and hence also observ*ER*-dependent). A "synthetic *a posteriori* truth" is a synthetic *a posteriori* statement that is true. A "synthetic *a posteriori* falsehood" is one that is false. A "synthetic *a posteriori* judgment" occurs when an observer decides, in an observation-dependent (empirical) way, that a synthetic (observer-dependent) statement is true (or false).

Definition 1.12. From the definitions already given, an "analytic *a priori*" or "*a priori* logical" statement is one whose truth follows from logic alone, and which is knowable prior to any *particular* experience. An "analytic *a priori* truth" is an analytic *a priori* statement that is true. An "analytic *a priori* falsehood" is an analytic *a priori* statement that is false. An "analytic *a priori* judgment" occurs when an observer decides, through logic alone, that an analytic statement is true (or false).

Definition 1.13. From the definitions already given, a "synthetic *a priori*" or "*a priori experiential"* statement is one that is experiential, but nonetheless knowable prior to any *particular* experience—*i.e.* observ*ER*-dependent, but not observa*TION*-dependent. A "synthetic *a priori* truth" is a synthetic *a priori* statement that is true. A "synthetic *a priori* falsehood" is one that is false. A "synthetic *a priori* judgment" occurs when an observer decides, in an observation-independent way, that a synthetic (observer-dependent) statement is true (or false).

Definition 1.14. From the definitions already given, an "analytic *a posteriori*" or "*a posteriori* logical" statement is one whose truth follows from logic alone, but which nonetheless is not knowable prior to any *particular* experience. An "analytic *a posteriori* truth" is an analytic *a posteriori* statement that is true. An "analytic *a posteriori* falsehood" is an analytic *a posteriori* statement that is false. An "analytic *a posteriori* judgment" occurs when an observer decides, in an observation-dependent way, that an analytic statement is true (or false).

It is tempting to take the analytic/synthetic distinction to be a metaphysical distinction, since it is fundamentally about what makes a statement true, while considering the *a priori*/*a posteriori* distinction to be more epistemological, since it is about how one comes to know that something

is true. While there is something to that, it can be confusing to make a hard distinction, just as it would be misguided to make a hard distinction between metaphysics and epistemology to begin with.

We can make another rough distinction by noting that any of these labels can be applied either to *(1)* statements and truths, or *(2)* judgments and knowledge—the former being a more metaphysical matter, and the latter more epistemological. Hence, a "synthetic *a priori*" truth is a statement that depends for its truth on experience (in general or in particular), but is knowable best (with most certainty) without appeal to *particular* experiences. However, it might still be possible to confirm such a truth empirically (*a posteriori*), which would count as a synthetic *a posteriori* judgment. The *statement* is still *a priori*, however, so long as an *a priori* demonstration is possible, in principle. In the meantime we assign the provisional label of "*a posteriori*", since this is what the statement currently is *for us*, in practice (given our current state of knowledge, which of course could always change).

Definition 1.15. If a statement is currently known *a posteriori*, and it is unknown whether an *a priori* demonstration is possible, we will refer to the statement as "provisionally *a posteriori*".

A perfect example of a provisionally *a posteriori* analytic statement is the Goldbach conjecture (that every even integer greater than 2 is the sum of two primes). This statement has never been proven analytically, but has to-date been empirically confirmed for the first four quintillion (4×10^{18}) even numbers greater than 2. As a result, many people have judged it to be (most probably) true. The judgment is *a posteriori*, even though the statement is purely analytic, and must be true or false by logic (or math or computation) alone. It is thus clearly possible to judge an analytic statement true *a posteriori*.

This does *not*, however, mean that the statement is analytic *a posteriori*—it may or may not be. Let us suppose that there is no possible proof for the Goldbach conjecture, and that there is no way to know whether it is true by *a priori* means, even in principle. This could well be the case, even though it is an analytic statement, since it might be that one would have to check the statement against all (the infinite number of) possible even integers to prove it true, which is clearly impossible. If this is indeed the case for the Goldbach conjecture, then it is truly an analytic *a posteriori* statement. On the other hand, this might *not* be the case: perhaps there is a finite analytic proof for the Goldbach conjecture, and no one has discovered it yet. If so, then it can actually be known with greatest certainty through *a priori* means, and the statement is thus actually analytic *a priori*. Since we currently have at hand no such analytic proof for the Goldbach conjecture—even though it is widely believed to be true for *a posteriori* reasons—all we can really say right now about the Goldbach conjecture is that it is an analytic statement of which we do not know whether it is *a priori* or *a*

posteriori, but that it is currently known only *a posteriori.* The most, therefore, that we can say is that it is analytic and provisionally *a posteriori.*

Definition 1.16. If a statement is analytic *a posteriori* (logically true but knowable only empirically), then we will refer to the statement as *"a priori* in principle only", or (along with all genuinely *a priori* statements) *"a priori* in principle" or "essentially *a priori.*" (Note that this means that all analytic *a posteriori* statements are essentially *a priori*—historically, in fact, it is usual to simply call them analytic and *a priori,* the former being taken to imply the latter. Note that Gödel's incompleteness theorem [93] can be interpreted, under the definitions given here, as a proof that there exist analytic *a posteriori* statements).

Definition 1.17. If a statement is *a priori,* but the required *a priori* demonstration is beyond practical reach, then we will refer to the statement as *"a posteriori* in practice only", or "practically *a posteriori*", or (along with all genuinely *a posteriori* statements) "*a posteriori* in practice".

Note that, while "*a posteriori* in practice only" and "*a priori* in principle only" are distinct, they will quite frequently be impossible to distinguish in practice. How could we know, for instance, which category the provisionally *a posteriori* Goldbach conjecture falls into (even assuming that it is true)? If an analytic proof is possible, and we just haven't thought of it yet, then the statement is straightforwardly *a priori*, and the *a posteriori* label is merely provisional. However, what if such a proof exists, but remains forever beyond human reach? If this is merely a practical matter, and such a proof is technically possible—even if humans had to devote all their money and time for millions of years into constructing such a proof—we would be compelled to admit that the conjecture is strictly *a priori,* not merely in principle, even though it remains *a posteriori* in practice. If, however, the proof is genuinely impossible for human beings to carry out, perhaps because it would require completing an infinitely long proof sequence, then the statement is strictly analytic *a posteriori*, but *a priori* in principle. But clearly, it may not always be possible to know which of these scenarios holds, since the proof is currently beyond our capabilities. To further confuse the issue, there is the possibility that the proof exists (it is possible in principle for *someone* to carry it out), but it is impossible in principle for *humans* to carry it out. Our existing definitions don't really help us in that case (but I would personally prefer to simply say that the statement is *a priori* in principle but *a posteriori for humans*).

As you can see, while it is important to understand the distinction between these various flavours of analyticity, more often than not, we will have to lump them all together as simply "*a priori* in principle", which happily applies to all analytic statements. The precise status of many statements may never be known, and may perhaps be unknowable.

Synthetic truths typically have to do with worlds and/or minds, since these are the two things we tend to take as given that nonetheless cannot be "proved", as they have an (at least apparent) resistance to analysis. We know our own mind, and the world around us, through direct experience, not logic. Nonetheless, some (we will call them "rationalists") believe (or at least hope) that it is possible, in principle, to render all knowledge as logical or mathematical; while others (we will call them "empiricists") believe that there is something irreducibly experiential about existence that can never be mathematized or analyzed. Still others (we will call them "mystics") take this even further, by claiming that there is knowledge that is neither logical nor experiential in nature.

Definition 1.18. By "rationalism", I will mean the view that all so-called *a posteriori* statements are ultimately analytic, being either *(1)* only provisionally *a posteriori*, or *(2) a priori* in principle. In both cases, we simply lack the more precise *a priori* formulation of the statement—in the former case, we just haven't figured out yet how to render the statement *a priori*, and in the latter case, there are fundamental limits to our ability to do so. (Rationalism thus tends to give metaphysics a primary role in physics.)

Definition 1.19. By "empiricism", I will mean the view that there exist synthetic *a posteriori* statements that are not *a priori*, even in principle. (Empiricism thus tends to reduce metaphysics to a supporting role in physics.)

Definition 1.20. By "logical positivism", I will mean the kind of empiricism that also claims that there exist no synthetic *a priori* statements, even in practice. (Logical positivism thus tends to go hand-in-hand with the complete rejection of metaphysics as a discipline, essentially giving it no role in physics at all.)

Definition 1.21. By "mysticism", I will mean the view that there exists knowledge that is neither synthetic nor analytic. (Mysticism thus tends to make metaphysics a completely separate discipline from physics.[3])

Definition 1.22. By "spiritualism", I will mean the kind of mysticism that also claims that there is synthetic *a posteriori* knowledge that is irreducibly mystical—its being ultimately caused by "spirit". (Spiritualism thus encourages the mystical view of metaphysics, rejecting the empiricist and rationalist views.)

Definition 1.23. By "materialism", I will mean the kind of empiricism that also claims that there is synthetic *a posteriori* knowledge that is irreducibly synthetic—its being ultimately caused by

[3]It is worth pointing out that in the popular press, the word "metaphysics" is frequently equated with this sort of mystical metaphysics, by defining it as the study of "that which is beyond physics," or somesuch. This is not the traditional meaning of the word, and not how the word is used in academia.

"matter" and/or "energy". (Materialism thus encourages the empiricist view of metaphysics, rejecting the mystical and rationalist views, including any kind of idealism.)

Definition 1.24. By "idealism", I will mean the view that synthetic *a posteriori* knowledge is ultimately knowledge of the objects of consciousness, whatever their cause or source. (Idealism thus tends to consider metaphysics to be the study of possibilities or ideas, and thus rejects empiricist views of metaphysics, including any kind of materialism.)

Definition 1.25. By "synthetic idealism", I will mean a kind of idealism that also claims that there is synthetic *a posteriori* knowledge that is irreducibly synthetic in nature. (Synthetic idealism thus tends to discourage a rationalist view of metaphysics.)

Definition 1.26. By "subjective idealism", I will mean a kind of synthetic idealism that also claims that synthetic *a posteriori* knowledge is always of something ultimately mental in nature. (Subjective idealism thus tends to encourage a mystical view of metaphysics.)

Definition 1.27. By "objective idealism", I will mean a kind of idealism that also claims that synthetic *a posteriori* knowledge is always of something ultimately non-mental in nature.

Definition 1.28. By "analytic idealism" (or "logical" or "mathematical" or "computational" idealism), I will mean a kind of objective idealism that also claims that synthetic *a posteriori* knowledge is always of something ultimately analytic in nature. (Analytic idealism thus tends to encourage a rationalist view of metaphysics.)

Definition 1.29. By "transcendental idealism", I will mean a kind of objective idealism that also claims that the objects of synthetic *a posteriori* knowledge are (*qua* objects in space and time) mental in nature, while remaining ultimately non-mental in nature (*qua* elements of reality).

Definition 1.30. By "transcendental synthetic idealism", I will mean a kind of transcendental idealism that is also synthetic, so that the objects of synthetic *a posteriori* knowledge are, *qua* objects in space and time, mental in nature, while remaining ultimately non-mental and irreducibly synthetic in nature.

Definition 1.31. By "transcendental analytic idealism", I will mean a kind of transcendental idealism that is also analytic, so that the objects of synthetic *a posteriori* knowledge are, *qua* objects in space and time, mental in nature, while remaining ultimately non-mental and analytic in nature.

Transcendental analytic idealism is thus a kind of synthesis of analytic and synthetic idealism, as well as rationalist and empiricist epistemologies. It operates functionally—for synthetic *a priori* reasons, within the domain of space and time—as a kind of synthetic idealism (*i.e.*, empirical objects

are, *qua* space-time objects, synthetic in nature). At the same time, it is more generally, and strictly speaking, an analytic idealism (*i.e.*, empirical objects are mathematical in nature *qua* elements of reality).

These "ism" definitions are intended to serve my own purposes in this dissertation, and are not intended to make everyone happy—there is no standard set of definitions for these terms, and I reveal my biases not only in my choice of definitions, but also in my choice of what to define and what to leave out. I have defined quite a few different varieties of idealism for a specific reason. Everettian quantum mechanics shares many features with idealism, and some of the most frequently voiced objections to it are essentially the same as age-old objections to idealism. In later chapters, we will deal with these objections, and the general issue of whether or not Everettianism is a kind of idealism (or whether it ought to be). While saving the details for later chapters, it will help to put things in perspective if I summarize now the general position I will be taking: in my opinion, there is no *a priori* reason to assume that Everettianism *has* to be idealistic or materialistic, empiricist or rationalist—and it is possible to address the Born rule debate without settling such issues (although Everett himself seems to have been an empiricist and materialist). However, given the resonance which Everettianism has with idealism, it is difficult to avoid the question as to whether Everettianism is any clearer, or more cogent. or more promising a framework, when interpreted in idealistic, rather than materialistic, terms. It has generally been developed, historically, in empiricist and materialist terms. I will argue, however, that the simplest and most elegant formulation of Everettianism, with most promise for addressing the Born rule (and other) objections, will strongly encourage us to consider the idea of an analytic and transcendental[4] idealism. In Ch. 6, I will argue that such a metaphysics can allow for in-practice-synthetic and in-theory-analytic space-time objects that are a mere tiny subset of all possible space-time objects—meaning that most of the "possible worlds" of idealism do not manage to qualify as "physical" or "material" in nature, due to information-theoretic considerations. I developed these ideas originally in [174], although I discovered later that they had already been developed by Robin Hanson in [102] (Hanson takes a very different approach, but the basic idea is the same). To use Hanson's metaphor, the most highly improbable worlds—the absurd or "maverick" worlds—are "mangled" out of existence by the more probable ones. Such a mangled-

[4]The reader may need to put aside any preconceptions of the word "transcendental", since it is widely used (or mis-used) in the popular press as a synonym for "mystical". I use it here to mean, not a metaphysical position, but an epistemological method, which may be employed by the mystic or the rationalist or the empiricist. It more or less refers to any means of acquiring knowledge by moving from one's *immediate experience* to that which transcends experience, but is necessary in order for that experience to exist in the first place. This method has its roots in Descartes' [68, 67, 69] famous credo "Cogito Ergo Sum" ("I think, therefore I am"). Possibly all metaphysics is transcendental, strictly speaking, but we tend to use the word only for those metaphysical systems that make the transition from immediate experience explicit, and pay it some systematic attention. The "transcendent realm" that is thus implied by immediate experience *could* be some mystical realm of spirit, but could also be the equally ineffable and indefinable corporeal realm of "matter". Or, it could be the definable and effable realm of mathematics. Each of these realms may be seen by their adherents as implied by our experience, while at the same time transcending it.

worlds view allows for an idealism that acts as a kind of operational materialism, negating most of the ethical and existential worries that people tend to have over idealism in quantum mechanics (not to mention, more strongly motivating the rejection of world/observer-counting needed to respond to the Born rule objection). I will return to mangled worlds later, as the basic idea is absolutely crucial to making an algorithmic interpretation of quantum mechanics workable.

Putting aside the possibility of mystical knowledge, which is beyond the concerns of this dissertation, our definitions thus far still leave room for the possibility that a statement might be neither strictly *a priori* nor strictly *a posteriori,* since it might be possible to judge it by either means, with no more certainty in one case than the other. However, this is unlikely to ever happen, since if it is possible to know a statement *a priori*, one would not expect an *a posteriori* demonstration to yield greater certainty. For our purposes, then, we will adopt the general rule of thumb that *a priority* always takes precedence over *a posteriority*, when it is available. To say that a statement is *a posteriori* will generally mean that it is not possible to know the statement *a priori*.

Historically, it was generally assumed that analytic truths and statements are necessary and *a priori*, and that the very idea of an analytic *a posteriori* truth is nonsensical. Some writers [148, 122] have, however, given "analytic *a posteriori*" meaning in different ways than I have, so the reader should not take my usage as standard (it would be more standard to just say that the category is empty).

The truth of a synthetic *a priori* statement is dependent on the prior (innate) nature of the observer—the way the observer comes already "wired up", so to speak—but not on any particular (posterior or empirical) observations. The "observer" in "observer-dependent" does not have to mean, specifically, *this* particular observer (although it might); it could refer instead to a larger class of observers that *this* observer is a member of. We can imagine different kinds of synthetic *a priori* truths, based on the size and scope of this larger class (and we will discuss some of the possibilities later).

It might seem equally obvious that synthetic statements, being experiential, are always *a posteriori* and contingent, and never *a priori* or necessary. Since these two categories would cover both logic and empirical observation, many would see this as exhaustive. However, Kant [109] argued, and indeed his philosophy was largely based on, the contention that synthetic *a priori* statements *are* possible. Take, for instance, the Pythagorean theorem (that the square of the longest side of a right triangle is equal to the sum of the squares of the other two sides). Kant argued that geometrical truths such as this are not analytic, but synthetic *a priori.* A proof of this theorem is not a purely logical affair, according to Kant, but depends on one's intuition of space and time, and is thus in some way experiential. Logically, it could have been otherwise. In other words, to convince yourself

of its truth, you need to imagine a triangle in space, and in so doing you make assumptions about space that are not logically justified. For instance, you assume space is flat (Euclidean), whereas space may be hyperbolic or spherical, for all you know. Humans generally take three-dimensional Euclidean space as an *a priori* (prior to *particular* experiences) because it is a necessary condition for having spatial experiences—in other words, we are innately built to picture space this way, even though it logically could be otherwise.

A crucial question for many synthetic *a priori* statements is whether they can be rendered analytic by giving them the appropriate context. For instance, suppose we agree with Kant that the Pythagorean theorem, as stated, is synthetic *a priori*, since it depends on innate human spatial and temporal intuitions. There are a couple of ways we could try to render this statement analytic (when I say "render" it analytic, I mean come up with a slightly different but related statement, perhaps more precisely phrased, that *is* analytic). We could simply adopt the stance of analytic geometry, which would have us reduce the statement to an algebraic one without specific geometrical content. However, then there would be no justification for preferencing the Euclidean features of space that we thought made the theorem true. So, generally, what is actually done is that the *a priori* synthetic qualities of space and time are axiomatized, and then the theorem is proved *from* this synthetic base, as a purely analytic proof. This, of course, does not make the truth an analytic one, *unless* we take the axioms to be purely hypothetical. For this reason, it is sometimes ambiguous whether a given statement or theorem is synthetic or analytic. We need to be specific about what "statement" we are asking the question about, if we care to decide the matter. For instance, one could argue that, of the following two almost-identical statements,

"If Euclid's Five Postulates are true, then his 47th Proposition follows"

and

"Euclid's 47th Proposition is true,"

the former is analytic, while the latter is synthetic (we put aside issues concerning whether Euclid's postulates are truly sufficient to axiomatize geometry—assume for argument's sake that they are.)

There are thus many different flavours of "synthetic" and "synthetic *a priori*".

Definition 1.32. For instance, we can define the following three different (but not exhaustive) kinds of observation-independent, observer-dependent (synthetic *a priori*) statements:

1. "Persona-synthetic *a priori*" or "person-dependent" statement: an experiential statement that depends on a particular observer's innate nature, but not on any particular experiences that observer has had.

2. "Anthro-synthetic *a priori"* or "human-dependent" statement: an experiential statement that depends on innate human nature, but not on any particular experiences that any humans have had.

3. "Cogito-synthetic *a priori*" or "consciousness-dependent" statement: an experiential statement that depends on the inherent nature of consciousness, but not on any particular experiences that any conscious beings have had.

That Euclidean geometry is synthetic *a priori,* Kant clearly has a case to make (although we will not try to settle its correctness here[5]). However, even if Kant is right, while Euclid's proof for his 47th Proposition may be both person-dependent and human-dependent, it is not necessarily consciousness-dependent, as it is feasible that there might be possible conscious beings (it is irrelevant whether such beings actually exist) who find it natural to think in terms of a five-dimensional hyperbolic space, and for whom the notion of a three-dimensional Euclidean space is quite alien. It could even be possible that such beings might actually live in a three-dimensional Euclidean universe—after all, humans have an innate intuitive preference for three-dimensional Euclidean space, even though our universe is actually *not* Euclidean. In other words, synthetic *a priori* truths are perspectival in nature—tied to a point-of-view, and in some sense always at least partly *about* that point-of-view—they are not automatically generalizable to some corresponding *a posteriori* or analytic statement, by somehow "detaching" them from their experiential context. Thus, it is perfectly possible for a human to *correctly* state that the 47^{th} Proposition is true, as a synthetic *a priori* truth, yet false as an *a posteriori* statement. Further, if the statement is formulated so as to make its own experiential context explicit, it may not even be possible to interpret it in an *a posteriori* context, without rewriting the statement.

The fact that we know *a posteriori* that our universe is *not* Euclidean is a lesson not to get caught in the trap of thinking that any given *a priori* synthetic truth necessarily generalizes beyond whatever experiential context it was framed in (whether that means generalizing it from a synthetic domain to an analytic one, or just from one synthetic domain to another one). That we innately experience geometry in Euclidean terms does not automatically mean that Euclidean geometry is the simplest kind of geometry that *explains* these experiences, *a posteriori.* On the other hand, the fact that some other geometry is known *a posteriori* as a well-confirmed physical law, does not necessarily mean that its truth is entirely *a posteriori.* It could also be, that were we to learn more about perception and experiences, we would understand that this non-Euclidean geometry was an *a priori* necessary condition for any such experiences. Such a truth would be synthetic *a priori*, even if, in practice, we had *judged* it to be true *a posteriori.*

As you can see from the preceding discussion, it can be very tricky to apply the analytic-synthetic

[5]For my views on Kant's philosophy of geometry, see [173].

distinction, especially when the *a priori* synthetic category is permitted. One reason is that we never actually prove something to ourselves in an absolutely *purely* logical manner. Proofs always involve some kind of synthetic, experiential cognitive structures, that we build in order to be able to make the logical inferences that we do—we cannot do "pure logic" in our heads. Our natures demand some kind of synthetic dressing, and it can be very obscure how much of our analytic proofs or constructions actually depend on this dressing. Demonstrating a proof in predicate logic, adding $2 + 4$, and writing a computer program are all "analytic", according to the definitions I have laid out above. Yet, someone can always come along and try to make a case that the substance of one's analysis is really synthetic—being highly dependent on the experiential context of the language used to perform the analysis—and it can be difficult to defend against this kind of de-construction.

The theory of computation, however, has produced a number of results that give us reason to think that analysis is not somehow entirely deconstructible into synthesis. One cannot "dissolve away" objective, analytic content by the trick of translating into multiple different languages, showing that there is nothing left that survives all such possible translations. Once one manages to ignore and put aside all the particular notational and intuitional supporting structures of even a huge untold number of formulations in different languages, there is still an analytic core that will remain necessary, objective and non-experiential. If one could show, for example, that *(1)* a predicate logic proof, *(2)* the adding of 2 and 4, and *(3)* the writing of a particular computer program, were all formally equivalent to each other, one could then be confident that what the three formulations have in common, after translation, is much closer to the true analytic content of all three, than what can be gleaned from any of the individual formulations alone. We can show this formal equivalence by translating back and forth between the different languages (predicate logic, arithmetic and the computer programming language). We will examine these ideas in more detail in Ch. 4; for now, it will suffice to note that while there may always be room for debate as to where analytic and synthetic content begins and ends, we will accept that there *are* analytic truths that are entirely independent of experience.

It is crucial to note that, just from the definitions given, there is no reason to suppose that there are synthetic *a priori* truths that are un-analyzable—that cannot be made into analytic truths by broadening the context, and appropriately analyzing the observer himself. However the analytic truth *qua* analytic, will not depend on the fact that it happens to describe an observer with a certain nature *qua* observer. The moment we make a statement's truth (or a computer program's output) depend on the statement/program's observer-ness, then its truth value or output is automatically synthetic, even though the statement or program itself is still entirely analytic. But, then, this is no different from any "outputs" or "truth values" of any formal system, which are never analytic, anyway,

but always synthetic. Truth values (qua truth values) are synthetic, since a formal, logical inferencing system does not consider anything special about the "true" or "T" symbol, that makes it correspond to truth, as opposed to falsehood. Even the idea of "contradiction" is largely synthetic, its formal significance being something more like "comprehensiveness" (a system that derives a contradiction can derive *any* proposition—but there is nothing *formally* or analytically contradictory about that; it only becomes a contradiction when we apply meaning to the symbols from outside the system). Likewise, inputs and outputs of programs are synthetic. Formally, a program just runs mechanically on its way, calculating. It is a matter of perspective to point to one "variable", or other internal state within the program, and baptize it as the system's "output".

Hence, if it were possible to fully analyze a synthetic *a priori* truth, it would still not be appropriate to talk about it as being "analytic synthetic," since it can only be made more analytic to the degree that we choose to drop the synthetic meaning we had previously ascribed to it. However, the fact that "analytic" and "synthetic" are, as such, opposing categories does *not* mean that acceptance of the synthetic *a priori* implies a belief that consciousness or observer-ness is unanalyzable and could never be fully captured in a logico-mathematical system, or computer program. Whether or not consciousness can be so analyzed is an entirely separate question from whether there are synthetic *a priori* truths. It should be noted, for our purposes, however, that the Strong AI position (which Everett essentially takes as an assumption) *does* assume an answer to this question; it assumes, in fact, that synthetic *a priori* truths *are* analyzable. Under this assumption, then—and only under this assumption—any complete explication of a particular synthetic *a priori* truth would necessarily contain a complete working analytic *model* of the observer within it. Speaking in terms of computer programs, this would mean a working simulation of the observer, including all of his neural and brain states that give rise to his consciousness. Such an analysis could be considered as nothing more than an analytic computational process. But, to the extent that our statements about this analytic process depend on the isolation within it of a consciousness, to that extent it must be considered to be synthetic.

Let us now return to the issue of Everettian probabilities—recall that I introduced the analytic-synthetic distinction, in the first place, with the promise that it would clarify the problem of the objectivity of Everettian probabilities. These probabilities *seem* quite strongly objective, since they are what they are no matter how much or how little we know, or what we believe, about the situation we are in. At the same time, they seem strongly subjective, since they are entirely experiential, being matters of perspective, rather than objective features of an external world. Our synthetic-analytic framework solves this apparent semantic conundrum.

First of all, we are not concerned here about how we know—or whether we know—that quan-

tum theory is true; nor about the truth of the outcomes of particular observations. We assume wavefunction realism here, and are asking only about the nature of our calculation of a probability *from* a given wavefunction state. This calculation (a kind of judgment) *is* strongly objective, since the result is the same, regardless of what beliefs or knowledge the observer has concerning it. At the same time, it is a synthetic judgment, because it is observer-dependent, since one cannot even calculate a probability without splitting the system into "observer" + "rest-of-the-system"—and, as we have seen, each of these subsystems does not even have a state of its own. More specifically, if Everett is right, it should be an *a priori* synthetic judgment, since Everett's wavefunction realism demands that the probability can be computed *a priori* from the wavefunction alone.

To be still more specific, it is a persona-synthetic *a priori* truth, since it depends on the personal identity of *this* individual observer—let's say her name is Liz, and Liz has just observed the "dead" result of a cat measurement. Now, analytically—meaning just from the logic and mathematics of the quantum wavefunction—one cannot determine a probability for this result, even if one knows the pre-measurement quantum state exactly. The reason is that, in order to calculate an *a priori* probability for Liz, we need to do *all* of the following (at least):

1. Factor the wavefunction ("divide it up") in such a way that Liz and the cat appear as two of its subsystems.
2. Show for the post-measurement state, how to express it as a superposition such that each "branch" is a coherent continuation of pre-measurement Liz's personal identity.
3. Determine what the states (and their amplitudes) are for the cat subsystem, *relative to* each branch for Liz.
4. From this information, calculate a probability.

Even assuming step 4 can be made entirely analytic, the resulting probability depends on the nature of Liz's identity as a conscious person, because of steps 1-3. Let's say it were actually possible for Liz to observe a dead/alive cat, given the nature of her perceptual equipment (and nothing about the wavefunction bans this possibility)—then step 2 produces only a single branch with a result that is 100% certain. This is clearly different from the case where Liz's perceptual equipment is only capable of perceiving a single result at one time, in which case we need to analyze the wavefunction in terms of two branches (dead and alive).

Hence, the probability calculation is observ*ER*-dependent, but not observa*TION*-dependent, since no particular observation needs to be made to do this calculation (presuming again Everett is right that quantum probabilities are *a priori*). In other words, Everettian quantum probabilities, if *a priori,* are persona-synthetic *a priori,* and strongly objective. Because they are synthetic, even though they are objective, one cannot deduce them analytically from the wavefunction. This is why,

unless the nature of the observer and the observation process (a.k.a., measurement) is itself fully analyzed in terms of the wavefunction (something we do not generally expect from physical theories, even when we presume it to be possible in principle), we cannot expect a strict mathematical proof of the Born rule. This doesn't preclude such a proof, but it gives us reason *not* to demand such a proof, the way the Born rule objectors do, since it is entirely possible such a proof would require a complete analytic model of human cognition.

1.7 The Synthetic Unity of Consciousness

It is possible to further hone the kind of synthetic truth a quantum probability calculation falls under, as it can be made even more specific than persona-synthetic. What is the reason that, after a measurement is made, we can make a determination as to how many branches there are in the post-measurement wavefunction state? We have already said that after the cat measurement, we analyze the state into two terms because this is the analysis that corresponds with the post-measurement consciousnesses that are mutually incompatible with each other, but all compatible with being continuations of the consciousness of the observer. This is a very special case of persona-synthetic *a priori* truth that relates specifically to the *unity* of an individual person's conscious experience.

Definition 1.33. A "system" is something "analyzable" (capable of being completely described by analytic statements) that is capable of having a state.

Definition 1.34. A "subsystem" is a part of a system that is not necessarily capable of having a state, but that can be said to have n states, each one relative to ("correlated with") specific states of other subsystems of the same system. (The degenerate case of $n = 1$ is permitted, and corresponds to a subsystem that has a specific state.)

Definition 1.35. "Synthetic unity" is the principle that, when a statement or truth is taken to be from a particular person's point-of-view or perspective ("the centre"), then there must be an analyzable system ("the universe") that can be analyzed into subsystems, such that there is at least one subsystem ("the observer") that has at least one relative state in the analysis that includes ("unifies") all information that is required for describing the consciousness of the centre. All other information in the analysis of the system (that is not required for describing the consciousness of the centre) must be isolated in another subsystem ("the environment"), which may itself be analyzed into further subsystems ("environmental subsystems").

The above may read as if it is about correlation in quantum systems, but there is nothing here that goes beyond applying our definition of synthetic *a priori* to the notion of analyzable observers in an analyzable world. If the whole system is analyzable (as Everett assumes) then we cannot

be allowed to simply "baptize" or "bless" certain subsystems or their states as being "*the* physical observer". Rather, observers and environments must be defined as arbitrary analytic subsystems—"arbitrary" because what baptizes a subsystem as an observer is its nature as a unitary consciousness, not some mysterious and undefined essence of "physical-ness". The preference for a "consciousness-based" analysis is therefore synthetic, not analytic, but without this preference, we cannot even talk about observations, so it is ultimately not really optional.

Likewise, the reference to "relative states" above is not an appeal to Everett, but a requirement for *any* system (quantum, classical or otherwise) that adheres to the same general philosophical assumptions as Everett's (realism with respect to the system as a whole, and a mechanistic view of observation/measurement). For imagine we performed an analysis of the overall system—say, as a computer program—and found that there was a definable subsystem (defined, perhaps, by a grouping of program variables) that contained all and no more than the information required to describe pre-observation Liz. Imagine further that this particular system/program contains only one subsystem with one state that can possibly be described as the person "Liz" prior to observation X. There is clearly no guarantee how many subsystem states such a system may contain that could be said to describe Liz just *after* observation X. There may be none (in which case, we would have to say the observation killed Liz). There may only be one, as expected in a classical world. Or, the computer program may generate multiple such states, each incompatible with the other (meaning they cannot be subsumed into the same consciousness). Since there is nothing inherent about any analytic system to prevent it from generating such multiple incompatible conscious states, we have to assume the possibility in general, for any *a priori* principle of synthetic unity. Whether or not such relative states are realized in actual physical systems is an *a posteriori* question, but we cannot assume either way, if we are to define an *a priori* probability measure for such systems (and the whole Everettian probability debate hinges on the identity of the correct *a priori* measure).

What if the system *does* contain multiple incompatible conscious states? What is the relationship between these states and their environment? Once again, recall that the system itself is entirely analytic, but we are constrained *a priori* to have a preference for putting the analysis of this system into a particular synthetic form (one that separates observer from environment). If there is more than one incompatible mental state described that yields the same person observing incompatible results of an observation, there is nothing to magically baptize certain variables in the environment as the ones that are "physically" under observation. The baptism must emerge *a priori* from our assumptions of analyticity of the system, and synthetic unity of consciousness.

Imagine, again, that our analysis is a computer program, imagine that one subroutine of this program is "Liz" and another is "the environment", and that we are compelled (by synthetic unity)

to consider "Liz" a subsystem (even though this is analytically arbitrary). Now imagine that the environment can be further analyzed into sub-subsystems, that these are sub-subroutines of the main program, and that one of them (call it "Cat") has multiple states (call them "dead" and "alive") that we wish to "match up" with the different incompatible Liz states (call them "seeing-dead-cat" and "seeing-live-cat"). Once again, there is no justification for a mysterious "physical" baptism that matches these up, by fiat. What would justify matching them up? Why can we not say that Liz's seeing-dead-cat experience is "about" the live state of the cat subsystem? Without the ability to correlate Liz's mental states with environmental states there is no justification for saying that Liz is really "in" an environment at all. We will call this the "correlation problem".

Keep in mind, again, that these are not quantum correlations or quantum subsystems we are talking about. The "system" here could be *any* analytically describable system with emergent mental states. Recall that the Born rule objectors are asking us to take the pure analytic description of the wavefunction—assuming with Everett that it contains emergent conscious mental states—and decide *a priori*, with no empirical input, how to calculate probabilities from the point of view of these emergent observers. We are taking the approach here that we need first to decide how to do this, *in general*, for an arbitrary analytic system with emergent conscious states, before we decide how to do it for the quantum wavefunction. Our arbitrary analytic system is not "physical" (and we can't require it to be physical if we wish to develop an *a priori* probability rule for it, since the whole concept of the "physical" or "material" is an entirely *a posteriori*, synthetic notion, and cannot even be defined analytically).

In Ch. 6, I will attempt to show how the correlation problem can be solved algorithmically, by using information theory. In this introduction, I will simply outline the main idea. For each environmental state that overlaps (shares "mutual information") with an observer state, the two states are said to be "correlated" or to exist "relative to" each other (and the observer and environmental subsystem are said to be "correlated)".

Let's start by looking at an example. Imagine an analytic system written in the popular BASIC computer programming language. This analytic system can be shown to produce a conscious observer Liz, in a state we will call "Liz-ready-to-measure-cat". The same analytic system generates *two* other conscious states that remember having *just* been "Liz-ready-to-measure-cat". We say they are "continuers" of Liz. Generally, they do not literally have to *consciously* remember being the ready Liz, but for now we will think of it that way. Neither of these conscious states are aware of the other. One we will call "Liz-seeing-living-cat", the other "Liz-seeing-dead-cat". In addition there are two other states in Liz's environment, neither of which produce any consciousness consistent with Liz's personal identity—one of these we will call "live-cat" and the other "dead-cat".

The analytic correlation problem is that, no matter how much we think that "Liz-seeing-living-cat" *seems* correlated with "live-cat", and "Liz-seeing-dead-cat" seems correlated with "dead-cat", there is no way analytically to justify this. This is just a formal system producing states, so there is no "physical wall" between two parts of the running program, making them two separate "worlds" or some-such. The reason algorithmic probability can help us out here is that it provides a measure (of information content) that allows us to correlate states within a formal system, in a systematic way. The "information content" or "entropy" of a state is, roughly speaking, the average number of bits required to completely describe that state. It turns out that this is nearly the same thing as the smallest number of bits, since smaller programs by far outweigh longer programs, in terms of probabilities. The "mutual information" of two different states is the informational "overlap" between them, or the number of bits of information they share. In §4.3.7, we will see that there is a precise formula for this measure, and that it relates directly to probability. The smaller the number of bits we can describe a state with, the more probable it is. The higher the percentage overlap between two states, the more highly correlated they are, and the more probable either one will be if we are already "given" the other.

Returning to the analytic correlation problem, imagine that the shortest possible program that can produce "Liz-ready" as its output, also produces "Liz-seeing-living-cat" and "Liz-seeing-dead-cat" as internal (non-output) states. Further, assume that the shortest program that outputs "Liz-seeing-living-cat" requires just one bit more than the ready program, as does the shortest program that can output "Liz-seeing-dead-cat". They will require at least *something* more than the ready program, since now we must distinguish which of these two "seeing-cat" states is the output. Note that the "seeing-dead" and "seeing-live" states will have the same length of shortest program, and will be equally probable (there will be a 50-50 chance of "seeing-live-cat").

Now we measure the mutual information between pairs of these states. If the "seeing" states are each highly correlated with the "ready" state (having high mutual information with it), and they both have memory records that could consistently be said to (in some sense) "remember" being in the ready state—to continue it—then we will consider that the "ready" state is in the past of *both* of the continuer "seeing" states. If the continuer states are not synthetically consistent with each other (representing incompatible mental states that could not exist in the same mind at the same time), then we say that the past "ready" world has now "split" into two future worlds. Now we measure the mutual information between the "seeing-live-cat" state and each of the "live-cat" and "dead-cat" states. Further, if the mutual information that "seeing-live-cat" has with "live-cat" is higher than it has with "dead-cat", we conclude that "seeing-live-cat" is more strongly correlated with "live-cat" than with "dead-cat", and this gives us more justification for saying that the two "live" states are in

the "same world" (and similarly for the two "dead" states). We have solved the analytic correlation problem, by recognizing *(1)* that consistent memory traces in the observers mind are a *synthetic a priori* requirement for "world-ness" and for the ideas of past and future, and that *(2)* correlation, defined in terms of mutual information, gives us an *a priori* analytic tool for further delineating one world from another.

But what about the question of branch-counting versus amplitude-counting? Many Born-rule objectors and Everettians alike would like to work under the assumption that some form of branch-counting is the logical *a priori* method for computing probabilities, given an arbitrary analytic system with emergent observers (since they claim this to be *a priori* true for the wavefunction, why would they claim any different, for more general analytic systems?). However, algorithmic probability tells us that different mental states, being synthetic, may have very different analytic probabilities, determined by program length. Hence, in the quantum domain, this gives us reason to think that perhaps amplitudes might actually be "program counts" of a sort, counting the number of (analytic) programs that can consistently produce a particular (synthetic) phenomenon. If so, not all "worlds" defined in terms of mutual information and consistent memory records, are necessarily equally probable. Thus, world-counting cannot be assumed if quantum probabilities turn out to be algorithmic.

On the other hand, in our example, it might seem that "seeing-dead-cat" and "seeing-live-cat" must surely each require *exactly one more bit* than does "ready", since both conscious states are actually produced by the "ready" program, and therefore to create a program that outputs one or the other, all I need do is use the "ready" program, and add a single bit to distinguish which of the two "seeing" states I wish to output. Hence, the two "seeing" states *must* be equally probable.

However, this does not follow. It could be that there were *two* different programs that were able to output "Liz-ready", one of which was at least somewhat longer than the other. The shorter one will (mostly) dictate the probability of "Liz-ready", according to algorithmic probability theory. Assuming that it is the shorter one that produces "Liz-seeing-dead-cat", it follows that "Liz-seeing-dead-cat" will be a *higher* probability state than "Liz-seeing-live-cat". So the fact that there were exactly two outcomes, both highly correlated with "ready", and consistent with being its future state, does not after all mean that they will be equally probable.

The future states, however, will always be *at least* a little bit less probable than the past states. Even the most likely outcome of the observation—and even though its program will contribute the most to the probability of the past state—will still be less likely than the past state itself, since there will be at least a little bit fewer programs that output it than output the past state. This means, that if the mutual information method for delineating worlds is correct, future states will

always have higher information content (or entropy, as it is also called) than past states, so long as the observer and his world are splitting (and even if there is no split, the information/entropy will still never actually decrease). This gives us a very natural *a priori* formulation of the second law of thermodynamics.

Delineation of worlds and observers via mutual information and consistent continuers is presented here as an *a priori* method, and it has to be if we are to eventually apply it to the wavefunction to address the Born rule objectors. But is it truly *a priori*? I have assumed a measure (mutual information), and a notion of what it means for an observer to be a coherent conscious entity separate from its environment. But such criteria are, partially, synthetic, which is why some people will be loathe to call them "*a priori*" (although it should be mentioned that most Born rule objectors seem to invoke world counting as superior to amplitude counting, and it is certainly no less synthetic). The argument I have tried to present, and which I will flesh out in later chapters, is that an *a priori* measure *must* involve at least some synthetic considerations, and therefore we cannot ever have a fully analytic proof—the very idea does not make sense, since we are asking about the probability of an *experience* happening, which is an inherently synthetically defined entity. And even if synthetic concepts can be subsumed under analytic, by a complete analysis of consciousness and perception, no one thinks it is reasonable to demand a proof of *that* in order to choose the best interpretation of quantum mechanics. But short of this, one cannot ever *analytically prove* a probability rule. This parallels problems in the interpretation of probability itself. There are numerous competing schools of thought on the foundations and interpretation of probability, just as there are for quantum mechanics. This is so, even though there is a fully analytic theory of probabilities that everyone (mostly) agrees on. The problem is that probabilities need to be interpreted when we apply them to real actions and observations—or even imagined and highly unrealistic actions and observations—since this means applying the uncontroversial analytic theory in a synthetic context.

Nevertheless, it is not realistic to suggest that an *a priori* derivation of the probability rule should not be possible, so long as we recognize the existence of synthetic *a priori* truths. Once we admit that some things just cannot be otherwise *for us*—even though they are not logically invalid—simply because we are what we are, as thinking conscious beings, then it has to be admitted that some such *a priori* assumptions are valid. But it also means we need to be careful in making those assumptions, and only appeal to things that can be argued to truly be necessary. Simply stating by fiat that all worlds or observers or outcomes must be equally probable is in no way invoking the synthetic *a priori*. At best, it is an appeal to merely a suggestive and apparent simplicity and elegance—which is fine as a tentative assumption, since science tentatively prefers the apparently simple and elegant all the time—however, it is not *a priori*, which is what is being demanded of the Everettians, due

to their claim that the interpretation follows from the (analytic) formalism.

The specific kind of synthetic *a priori* we used to solve the correlation problem was persona-synthetic, meaning that it depended on the innate nature of a particular person and their personal identity. However, in particular, we appealed to the *a priori* requirements for *maintaining the unity and coherence of a particular person's stream of consciousness*, via the synthetic connection between mental states and their continuers.

Definition 1.36. "Unitary-synthetic *a priori*" statements are persona-synthetic *a priori* statements whose truth depends on the unity of a particular observer's consciousness, but not on any particular experiences that observer has had.

The principle of appealing to the unitary-synthetic *a priori* in order to establish the relationship (or correlation) between a conscious state and a state of something in the world, is very much like Kant's [109] similarly named "synthetic unity of consciousness"[6], which leads to Kant's "transcendental idealism", in which the unity of physical objects is a reflection of the unity of the observer's consciousness. While this is indeed very similar to what I have described above for determining a probability rule, I will not develop these ideas here in an exclusively idealist context. Nonetheless, the question of idealism is important, and will be discussed in later chapters. The main point here is that the synthetic unity of consciousness is required in order to compute a probability *a priori* from a *purely* analytic structure, and it is an *a priori* derivation that is being asked for by the objectors. Of course, if we suppose materialism or physicalism to be true, and idealism false, then it is possible that the *physical* nature of a system provides the "wall" between subsystems that makes them separate, and the "wall" between worlds that correlates states within a world—and then, the appeal to synthetic *a priori* methods would not be required to solve the correlation problem. In this case, the quantum probability rule would be an *a posteriori* rule, and experiment should show a difference between the *a priori* version and empirical reality (Born rule objectors argue exactly this, by claiming that world-counting is the *a priori* method, which is falsified by experiment). It is also possible that materialism may be true, but that it does not affect the way probabilities are calculated—not by building a "wall", nor via any other physical influence—in which case we should expect that experiment will verify the *a priori* rule, just as it would if idealism were true.

Synthetic unity of consciousness is a similar concept to *personal identity*, but subtly different. We can argue whether "me now" is the same *person* as "me last year", just as we could argue whether ten different experiential observers in superposition are really the "same person" or different people. This all depends on how finely coarse-grained we want our concept of "person" to be, which is

[6] It is actually called the "synthetic unity of apperception" in most English translations, but the meaning of "apperception" is essentially the same as "consciousness", which is a more readily understood term.

highly arbitrary and subject to the requirements of a particular application. The fact that I have a particular immediate conscious experience, however, is not something I can re-define or change according to circumstance. The fact that the ten different individual observers in superposition are *different* consciousnesses is not something that can change depending on that person's idea of themselves, or beliefs. Nor is it something I can make go away by insisting that the ten *different* consciousnesses are in reality "one physical observer".

Here is the crux of this whole point: as we established in our discussion of subsystems, it is analytically arbitrary (and usually subjective) to split a system into subsystems. We usually make this split because the division into "subsystem 1" and "subsystem 2" reflects something that we *care* about (and caring is subjective). However, the split of a system into "my consciousness" and "everything else" is *not* arbitrary, and hence is objective in spite of being observer-dependent. While probability measures based on caring, fatness and synthetic unity are all synthetic, only the synthetic unity-based measure can claim to be objective or *a priori*.

Of course, even if we decide that synthetic unity must be appealed to, this does not tell us whether we should count worlds or amplitudes, or some other thing. It only lays the groundwork for why we are permitted, *a priori*, to partition the overall system and categorize its subsystems in the way that we do. Normally, in any application of probabilities where there is any pretense to the results being "objective", there is still a synthetic, and often entirely subjective, categorization of events into event classes. However, the actual *things* that we count are still observer-independent things. Let us call objects of this type "ontic entities".

Definition 1.37. For any class of objects, if the objects in that class are not dependent on observers—in particular or in general—for their existence *qua* objects, then they are "ontic entities". If they can be in some sense "counted" (by which we shall mean discretely counted, summed up, or continuously integrated), so that the resulting "count" (or sum or integration) is independent of the observer or point of view of any observers, in particular or in general, then the members of the class are "countable ontic entities".

Generally, I will mean countable ontic entities when I speak of "ontic entities" in this dissertation, since we will be generally assuming there is *some* way to compute objective probabilities, if we have complete knowledge about the ontic entities of a system. If there is not, and our ontic entities are not countable ontic entities, then the entire Everettian project to derive probabilities as *a priori* chances from the wavefunction would seem to be doomed.

As an example, suppose I repeatedly pick a marble from a bag, returning it to the bag after each pick. Assume the picking procedure is "random" (putting aside for now exactly what that means). Empirically, I observe a 70% chance of picking a red marble, and 30% chance of picking a

blue one. I therefore assume that there is a store of marbles in the bag, 70% of which are red, and 30% blue. The choice of "color" in categorizing our marbles was arbitrary and subjective—revealing not an objective preference in objective reality, but simply something that I personally care about. Nonetheless, my probabilities can still be otherwise objective, so long as I can make the case that *marbles are ontic entities.* If marbles are the countables, then I can say that my "random pick" from the bag was a "random pick of a marble", and my probability rule for marble picking has objective validity (*given* my arbitrary choice of categories).

The choice of "marble color" as the thing that matters is the kind of choice that, while in general subjective, might be "objectivized" by appealing to the synthetic unity of consciousness. But the choice of "marbles" as the ontic entities that are "picked" depends on what *really are* the things out there in objective reality, independent of any observer. If we were to find that "picking a marble" is not really what is happening, but rather there is only a single marble in the bag, and what is happening *in reality* is that this one marble turns a random color whenever we pull it out of the bag, then our ontic entities would actually be something more like "colors", and not marbles.

So a good (*a priori*) probability rule will take the analytic structure that is given (such as a wavefunction) and separate it into observers and correlated subsystems according to the synthetic unity of consciousness (perhaps using mutual information as a correlation measure, if we are employing algorithmic probability theory). However, the actual counts that are used to compute the probabilities (the "marbles" of the system) will be observer-independent countable ontic entities. Otherwise, we cannot claim the objectivity for our probability rule that we need in an Everettian project (and then we might as well dump synthetic unity and just choose fatness or caring).

For quantum mechanics, these entities *cannot* be worlds or observers or outcomes, any more than they can be caring or fatness, since none of these are observer-independent. The reality is that the wavefunction is ultimately composed of—you guessed it—amplitudes, not worlds or observers. Amplitudes are not merely relative to the observer (although *which* amplitudes we care about depends on the observer). They are an inherent part of the wavefunction, with an objective and analytic definition. Hence, they are a viable candidate to be a countable ontic entity.

Nonetheless, this still isn't a *proof* that we must count amplitudes, although it leaves us a little perplexed as to just what our alternatives are supposed to be—a perplexity that I will call the "What else?" response to the Born rule objection: if not worlds, observers or outcomes, and not amplitudes either, then just what exactly *are* we supposed to count?

1.8 The Problem of Interference

I hinted earlier that, even if we were to convince all the Born rule objectors that counting worlds (or observers or outcomes) has no *a priori* priority, that it would still likely be an uphill battle to convince them to count amplitudes. The reason is the presence of interference effects, which make amplitude counting seem highly counter-intuitive, and a definite difficult sell as an *a priori.* Keep in mind that nobody is actually denying amplitude-counting; the Born rule is part and parcel of quantum theory. The question instead is whether a rational person would naturally assume amplitude-counting, if presented with a many-worlds, wavefunction-based universe—simply as an intellectual exercise, perhaps, without any empirical support for it. I would suggest that the support for world-counting as an alternative would disappear overnight, if amplitudes could somehow be made to sum up nicely to probabilities, in a natural way, the way marbles do. Because it does not seem that they do, the popular thing is to assume that counting of worlds is the most natural thing, since worlds—even though synthetic in Everett's system—do seem to sum up nicely (at least on an intuitive level).

The problem with counting amplitudes is that quantum waves combine and interact with each other in a way that creates interference effects, both constructive and destructive. Imagine you throw a rock into a pond. A water wave ripples out in all directions away from the point where the rock entered the water. Now imagine you throw two rocks into the pond, a few feet apart. As before, a wave ripples out from each point of entry, but this time the two waves meet, creating a complex "interference pattern", as the two waves combine together. We get a higher amplitude wave where two crests meet (constructive interference), but a smaller amplitude wave where a crest and a trough meet (destructive interference). If the waves happen to have the same period, and the crest of one wave always corresponds to the crest of the other, then they just amplify each other, and we get another resulting wave identical to the first two, but with larger amplitudes—in this case, we say the waves are "in phase" with each other. If the crests do not match up, they are "out of phase". If they are perfectly out of phase, and the crests of one wave coincide with the troughs of the other, then we get destructive interference—everything cancels out and we get a perfectly flat wave with zero amplitude. If the waves are out of phase, but not perfectly so, we will get a sometimes complicated combination of constructive and destructive interference: an interference pattern.

Nothing in the classical world that we "count up" to get a quantity proportional to a probability behaves anything like this! Classically, we are counting "events". If there are two ways an event can happen, this should give a probability twice what we would get if only one of the possibilities

existed—this is, after all, what it means to "count up" possibilities.[7] But in quantum theory, because the quantities we are "counting up" have phases relative to each other—they are amplitudes of waves—there could perfectly well be two ways that the *same* event could happen, each with the same amplitude-squared probability measure, and yet because they sum or "count up" destructively, they cancel each other out, and the event ends up with zero probability of happening! The two different "ways" the event could happen might even be, in themselves, the exact same event. It is only their phase relationship *to each other* that determines the kind of interference that results.

In summary, it would be a lot easier to make the "what else?" argument fly, if it were clear how amplitudes could even be the countable quantity for a probability measure to begin with. In later chapters, I will suggest a natural *a priori* reason why we should expect to see interference if the countable ontic entities are algorithms or computer programs, based on synthetic unity and mutual information to solve the correlation problem. Based on this way of thinking about amplitudes—in terms of algorithms—it is just as natural and intuitive for event possibilities to cancel each other out, as it is for water waves to interfere with each other.

1.9 The Anthropic Principle

Algorithmic information theory is the branch of mathematics and computer science that measures the "information content" of physical systems, in terms of how many bits or bytes it takes to describe the system. For instance, just because I have a description of some physical system in a file on my computer, and that file takes up 10 Megabytes of space on my hard drive, does not mean that the information content of the system is 10 Megabytes. If I can compress the file down to only 1 Megabyte (as with the popular "Zip" program, for instance), then I must consider that the information content might be as low as 1 Megabyte (although it could possibly be even lower, since the zip algorithm might not be the most effective possible compression algorithm).

The anthropic principle (or "AP") is a guiding principle that reminds us that our very presence in the universe constrains the properties that the universe can have. It is not so much a theory as a philosophical principle that constrains theory formation. When combined with the MWI, it usually makes the point that if many universes co-exist, we are only going to find ourselves in universes that are consistent with life and consciousness—not because consciousness is somehow fundamental to the workings of the universe, but simply because we could not even be in a universe that did not have such consciousness-supporting features. Without many worlds, the suitability of the universe

[7] Counting is a little bit harder if we are dealing with an uncountable infinity of things, and much of modern "measure theory" is designed to deal with this complication, but the basic operation in these cases is "integration", which is still fundamentally a kind of (continuous, non-discrete) counting, or summing up, and does not fundamentally change anything here.

to life might be nothing more than raw coincidence (even granting the anthropic principle). With the addition of many worlds, however, the anthropic principle takes on an explanatory role. The reason the universe is suitable for life is that a huge plethora of universes actually exist, some small number of which are suitable for life. Our universe might, then, be considered a random pick from consciousness-containing universes (or, more precisely, a random pick from whatever ontic entities are generating these universes).

On a more local scale, each quantum measurement outcome might likewise be considered a random pick from possible outcomes that happen to continue the consciousness of that particular observer (or, more precisely, a random pick from the ontic entities that generate such continuers). Again, this is not because the consciousness of the observer is somehow playing a fundamental role in the mechanics of the situation, but rather because only certain observations are consistent with a single stream of consciousness in the first place.

We actually have two versions of the anthropic principle here, which differ in the scope of their respective use of the synthetic *a priori*:

1. *Persona-Anthropic Principle*: I should expect that the outcome of an observation will be a random pick from all ontic entities that continue my current consciousness forward with my personal identity intact (my "continuers").
2. *Cogito-Anthropic Principle*: I should expect that the world I am living in is a random pick from all possible consciousness-containing ontic entities.

The persona-anthropic principle seeks to make predictions about the future. The cogito-anthropic principle seeks to explain the nature of the world I already find myself in.

Under this view, the reason the half-live/half-dead cat outcome is never observed is not because the observer's consciousness collapses the wavefunction, nor is it because such a superposition is not mathematically a possible outcome of the observation, but only because such a superposed state *could not be consistently observed by a single consciousness.* The nature of consciousness, therefore deserves a role in determining the probabilities of outcomes, not because the conscious observer somehow selects the outcomes, but because the nature of consciousness places constraints on what outcomes are observable.

Note that these anthropic principles do not tell us how to compute objective probabilities. They tell us only that, in order to do so, we need to count ontic (objectively existing) entities—it does not tell us what those entities are. But, certainly, there is no good reason to presume that these entities are worlds or observers, just because those are the things that are synthetically "splitting" in the MWI.

1.10 An Algorithmic Anthropic Principle

Algorithmic information theory gives us an excellent candidate for "what to count" when computing probabilities. The more bits it takes to describe a system, the less probable it will be in a random pick from all possible systems. Applying this principle to quantum measurements, while employing an anthropic constraint, means that we compute the information content of all measurement outcomes consistent with a single stream of consciousness of an observer, and the more bits required for any given outcome, the less probable it will be. This means that our countable ontic entities are essentially abstract computer programs.

The justification for choosing programs as our objectively existing entities is simple: computation provides the only language that is fully analytically expressive: that can describe anything the human mind is capable of precisely describing. This does not mean that programs *are* the ontic entities of the universe, but it does mean that they are the rational choice. If the universe works in a way that is rationally explicable, then the ontic entities *must* be computer programs. It is possible that the ultimate things of the world are "material" or "spiritual" in nature, but these are things that are neither analytically definable nor describable, so if they are our ontic entities, it will not be possible to count them, and objective probabilities are likely to remain illusive. This makes abstract computer programs a very elegant and logical hypothesis for our countable ontic entities. The point here is not to establish a metaphysical dogma, but simply to put forward what appears to be the simplest and most rationally compelling initial hypothesis. Like any metaphysically-motivated hypothesis, it could turn out to be wrong when we actually come to empirically test it.

Assumption 1.38. ***Algorithmic Synthetic Unity (ASU):*** *the result of any observation, measurement, or other cognitive or perceptual act performed by a conscious state is a random pick from all of that state's continuer programs (all abstract computer programs that continue the consciousness of that state with the same personal identity intact).*

We could test the ASU hypothesis, in the context of quantum probabilities, if we had an *a priori* derivation of a probability rule based on these principles (since there is already an existing body of empirical evidence in support of a particular such rule, the Born rule). This would not, as some would have it, mean a formal derivation of a probability rule from the analytic content of quantum theory alone (*i.e.*, the wavefunction and its dynamical laws). If the result of a quantum measurement is a random pick from computational continuers, then any such derivation will have an *a priori* synthetic character. It will need to make at least some minimal assumptions about how consciousness and perception arises within the analytic wavefunction, and how this allows measurement to take place. So a purely formal proof is not to be expected, even if based on objective probabilities in a purely

formal ontology. My own attempt to justify the Born rule, from algorithmic assumptions, will take a form I prefer to call a "synthetic derivation". It will have gaps and unjustified assumptions that I hope are well-motivated, but which I will not claim to fully and adequately justify.

1.11 The Interpretation of Probability

All of this will have to wait until Ch. 6-8, however, since we will need first to establish a working interpretation of probability theory, without which no such synthetic derivations will be constructible, even in outline. I will follow neither the frequentist nor Bayesian paths (in fact, my interpretation of probability will have more in common with the classical view of probability that preceded frequentism). I will defend a "generative" interpretation of probabilities, in which we count out objectively existing generating entities to arrive at objective probabilities, but we perform this counting within the context of a particular observer. In order to do this, we need to decide what objective entities we wish to count, as the generative interpretation of probability will not itself take a stand on this. I will defend the idea that abstract computer programs are the appropriate *a priori* countable for probability theory, an idea originated by the founder of algorithmic information theory, Ray Solomonoff [203, 204]. This yields the algorithmic interpretation of probability, which I present as a special case of the generative interpretation. Applied to the problem of quantum probabilities, this should mean that the Born rule in some way expresses the probability of outcomes from randomly chosen computer programs—randomly chosen, that is, out of all those programs that generate the current observer. I will call this kind of counting "observer-algorithm counting". It does not count programs straight-out, but rather only programs that are algorithms for generating the observer in question. Nonetheless, since the objective entities it counts *are* themselves observer-independent, our resulting probability measure is objective, not subjective.

Solomonoff has made this task easier by providing us with a formula for calculating such probabilities, based on the length of the shortest program that generates the required output (in our case, that output is the state of the observer).[8] Unlike world-counting, this scheme in no way demands that worlds or branches or outcomes be equiprobable.

The problem, however, is that there is still potentially a version of the Born rule objection that holds in this case. If we assume that the wavefunction is a compression of a conscious state, then it

[8] This is not quite correct. Solomonoff's probability measure is actually based on the length of the *average* program that generates the output, rather than the shortest program. I will frequently talk in terms of the shortest program, however, since this allows us to conceptualize the measure as a form of data compression, an intuitive way to grasp the significance of Solomonoff's measure. But, as we will see in more detail in Ch. 4, Solomonoff's convergence theorem [205] shows that the shortest program is overwhelmingly the largest contributor to the average, and hence it is common-place in algorithmic information theory to talk as if the two measures are virtually (but not exactly) the same thing.

would appear that this "shortest program" is essentially a list of wave amplitudes. Yet, the number of bits used to represent each amplitude is the *same*—or at least we have no reason at present to assume otherwise—so algorithmic probability theory would seem to insist that the amplitude for each event is equiprobable, and we are back to world-counting.

However, I will argue in Ch. 6-8 that this new version of the Born rule objection misses a crucial lesson of algorithmic information theory. Such arguments assume that all of the possible algorithms for computing a conscious state are mutually exclusive. However, programs can have overlap, by having mutual information, even when the conscious states that they generate are mutually exclusive. In other words two programs can be *synthetically* mutually exclusive, while being *analytically* highly mutually dependent. This means they do not simply add up unless we eliminate the overlap. This overlap is essentially a form of "program interference" (of which wave-based or "periodic" interference is a special case). This explains why it is that two possibilities for the same event can destructively interfere, or even cancel each other out completely.

1.12 Cosmic Stability

Given that our world is stable and law-like, we will expect the algorithmic information content of a human consciousness to be a global, cosmic measure—not a local measure in terms of neurology or space-time coordinates. The reason for this is fairly simple. It is not the number of bits "in" our brains, but the number of bits in the average program that *generates* our consciousness, that matters. A priori, we cannot say whether or not this is the same as the number of bits in a perceptual or neurological description. It might alternatively correspond to the number of bits in a description of the entire universe in which the brain of the perceiver resides. Or, it might be a description at some level in between these extremes. The level of the description will not be a first principle. The description will be at whatever level it needs to be *in order to generate the conscious experience in question*, and to do so in the fewest bits possible.

If, indeed, a global description is needed to minimize program size, then, in order to recover the notion of a stable and lawlike environment, it follows that it will take *more* bits to describe one's brain state (at the local level) than to describe the entire universe, as I argue in [174]. Without this assumption, there would be no order or stability to the world. This seems counter-intuitive, since it seems to be saying that the universe, as a whole, is smaller than one's brain. However, since the universe *is,* according to cosmologists, unravelling from a simpler (lower entropy, or smaller program size) state, it is not really that unfeasible that it could be described in fewer bits than it takes to describe a human brain in isolation from its environment. In fact, if our cosmology could get to the

point that the information content (or entropy) of the early universe could be reasonably estimated, and it could be shown to require a higher number of bits than the local information content (or entropy) of a typical human mental state, the whole ASU framework would be falsified. Hence, in spite of being an "anthropic" measure (perhaps inviting accusations of unfalsifiability from some quarters), our measure could in principle be falsified one day, given enough understanding of the evolution of the early universe and the nature of human perception (enough to produce reasonable estimates of their relative information contents).

1.13 Synthetic *A Priori* Quantum Theory

Quantum mechanics is based on five fundamental postulates. In Ch. 8, I will argue that the essential content of these postulates can be inferred (*a priori* synthetically) from first principles, based on an algorithmic epistemology and interpretation of probability. This will be the latest iteration in my project to create a purely rationalist reconstruction of quantum mechanics, started in [169, 172] and continued in [174].

I will first present, in Ch. 6-7, a series of thought experiments and toy examples that will show that the general ideas of superposition, collapse/branching and probability interference, can be seen simply as the inevitable consequence of a more general attempt to derive a probability rule, from the perspective of observers emergent from *any* formal system. These toy examples will *not* be examples of quantum mechanics, but rather more general attempts to reason about observers and probability in formal systems. The fact that some of the more perplexing features of our empirical theory emerge naturally out of this process gives the overall algorithmic framework a great deal more credibility as a foundation for quantum mechanics.

In Ch. 8, I take this process one step further, by making one additional assumption concerning the optimal compression algorithm for conscious states. Clearly, we cannot definitively derive such an algorithm at the current time, but by looking at the nature of actual technological data compressors, and from informal arguments about what would be required to compress a conscious state, I will argue that a very reasonable educated guess can be made that the optimal consciousness compressor would be some form of a Fourier transform, which is a very common way to express many different kinds of data in terms of frequencies and waves. This "DFT hypothesis" gives us the essential analytic content of the quantum postulates, since the wavefunction, in its solved form, actually *is* a Fourier transform. Applying this to our existing arguments from the non-quantum toy examples yields a set of *a priori* principles and theorems that is very close to the full *a posteriori* quantum postulates. Close enough, I believe, that it constitutes a plausible, coherent and consistent interpretation of quantum mechanics.

I will argue that on this foundation it makes more sense than ever to simply accept Gleason's proof of the probability rule, augmented by an argument for noncontextuality in ASU. Details will have to wait until Ch. 8, but it is not hard to see in outline why ASU rules out contextuality. The only widely-considered alternative to the Born rule (branch-counting) clearly does not make sense within this interpretation, and any other non-Born rule one might imagine will still have to permit contextuality, allowing the probability of an outcome to depend on how the alternative outcomes are conceptualized, thereby violating the objective nature of probabilities in ASU. Thus, the very idea of contextual measurement simply does not make sense within the ASU framework.

1.14 Conclusion

Ultimately, the success of algorithmic synthetic unity will depend on the credibility of the claim that there are more bits in a local description of a human brain than in a global description of the universe. Without this, we have no reason to expect stability in the universe, and no reason not to observe crazy maverick worlds, rather than the orderly and structured universe we seem to have. Some may consider this claim to be extremely implausible, so much so that they will consider ASU already dead-in-the-water. How could it take more bits to describe just my brain than it takes to describe my entire world, which includes my brain as just one tiny, almost insignificant part? I will argue that what we know of modern cosmology and nonlinear dynamics (such as fractals and chaos theory) at least makes the claim more plausible than it might otherwise seem. I will also argue that the claim could constitute a source of future empirically falsifiable predictions, given that, as our understanding of the universe and of the human brain advance, we may actually be able, at some point, to offer reasonable estimates of the information content of each. If it turns out that the information content of the universe has more bits than that of the human brain alone, then ASU will be falsified.

2 Quantum Mechanics

2.1 Introduction

I will assume, in coming chapters, a basic understanding of Dirac bra-ket notation, and the essentials of quantum mechanics and quantum measurement. Readers unfamiliar with this material may consult any introduction to quantum theory that works in Dirac notation. This chapter is for those who do not have time to fill in the technical background in detail, but wish to have a more formal grasp of the measurement problem than the non-mathematical discussion provided in Ch. 1.

I do not pretend that what follows is a complete and adequate introduction to the topic, but I will briefly review the most essential material, which should make the remainder of the dissertation understandable in its essentials, provided the reader has a solid grounding in high school mathematics.

Those readers who do not have this background, and do not have the time to master this chapter, are invited to skip it or skim it, and read on. Much of the rest of the dissertation may still be understandable in its essentials (but expect to skip over some bits and pieces). Those readers who *do* have the required background can probably safely skip this chapter.

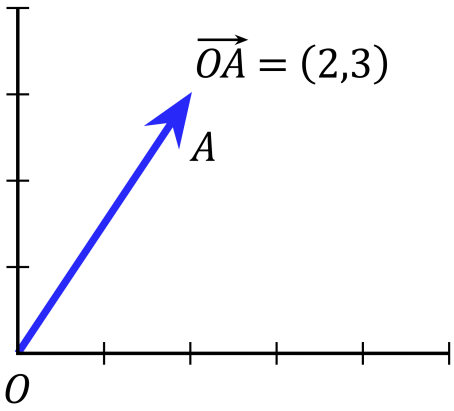
This chapter owes much to the pedagogy found in [198, 99, 189, 4, 53, 105], all of which are recommended for readers seeking a more complete introduction to this material.

2.2 Mathematical Background

2.2.1 Hilbert spaces

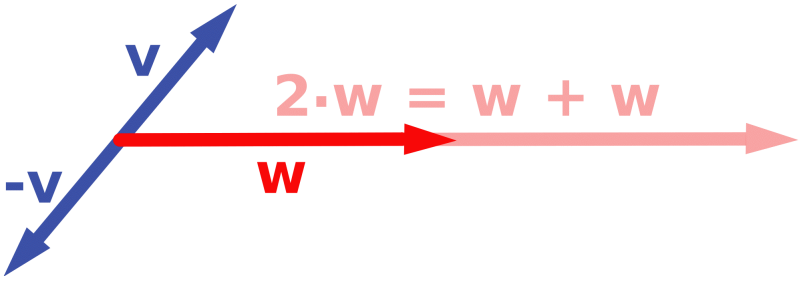
The state of a quantum system is represented as a "vector" $\mathbf{A}$ in a "complex Hilbert space" $\mathcal{H}_{\mathcal{A}}$. A Hilbert space can be thought of as a generalization of Euclidean three-dimensional space, in which vectors of arbitrary (even infinite) dimensions can be represented, and which has an "inner product" (an operator that takes two vectors and returns a scalar quantity).

The simplest, and most easily visualized, example is that of arrow vectors in Euclidean space (Fig. 2.1 shows an example).



Acdx (CC-SA 3.0); wikipedia.org/wiki/File:Position_vector.svg

Figure 2.1: An arrow vector in 2-D Euclidean space.



Jakob Scholbach (CC-SA 3.0); wikipedia.org/wiki/File:Scalar_multiplication.svg

Figure 2.2: Multiplication of a vector by a scalar.

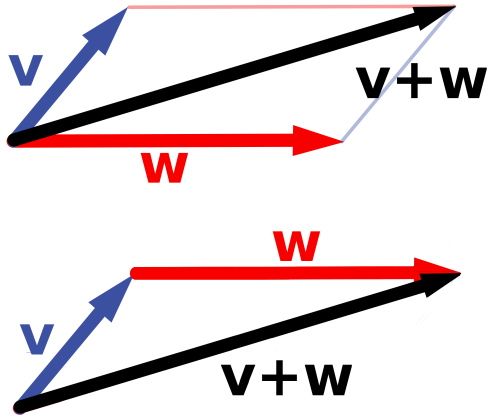
2.2.2 Multiplication by a scalar

One of the fundamental operations that can be performed on vectors is multiplication by a scalar (*i.e.* by a quantity with magnitude only, and not direction). This yields the expected result of producing a vector whose magnitude is simply multiplied by the scalar. So multiplying an arrow vector by the scalar 2 produces a vector that points in the same direction as the original arrow, but is twice as long (see Fig. 2.2).

2.2.3 Vector addition

Vectors can be "added" one to another. For arrow vectors, this means using the "head-to-tail" rule (place the tail of one vector on the head of the other, and connect the tail of the former to the head of the latter to form the new vector).

In Fig. 2.3, the new vector $\mathbf{v} + \mathbf{w}$ is "decomposable" into the two "component vectors" $\mathbf{v}$ and $\mathbf{w}$. Or, we can say that $\mathbf{v} + \mathbf{w}$ is a "linear combination" of the two vectors. And, since multiplying by a



From a drawing by Jakob Scholbach (CC-SA 3.0): wikipedia.org/wiki/File:Vector_addition3.svg

Figure 2.3: Vector addition

scalar affects only the magnitude of the vector, we know we can form $\mathbf{v}+\mathbf{w}$ as a linear combination of any multiples of the same components, such as $\frac{1}{2}\mathbf{v}$ and $2\mathbf{w}$:

$$\mathbf{v}+\mathbf{w} = 2\left(\frac{1}{2}\mathbf{v}\right) + \frac{1}{2}\left(2\mathbf{w}\right) \tag{2.1}$$

2.2.4 Orthonormal bases

Imagine that the vectors $\mathbf{v}$ and $\mathbf{w}$ in Fig. 2.3 can be combined in various linear combinations to form *any* arbitrary vector in our vector space. So *all* vectors in the space can be put into the form

$$c_1\mathbf{v} + c_2\mathbf{w} \tag{2.2}$$

for some scalar quantities c_1 and c_2.

We would then say that the vector set $\{\mathbf{v}, \mathbf{w}\}$ "spans" the vector space, because we can construct any vector in the space out of these two. This set is thus said to be a "basis" for the vector space, but only so long as it is "linearly independent" (so that no vector in the set can be constructed out of a linear combination of any of the others). Linearly dependent spanning sets are not considered appropriate bases because their elements are to some extend redundant. In addition, most convenient bases—barring some compelling reason to choose otherwise—are those whose elements are all orthogonal to each other, with each of unit length (= 1). The standard basis for 2-D Euclidean vectors over the real numbers is thus the orthonormal basis

$$\{(1,0),(0,1)\} \tag{2.3}$$

So the vector $(-2,1)$ could be said to be the linear combination:

$$-2(1,0) + 1(0,1) \tag{2.4}$$

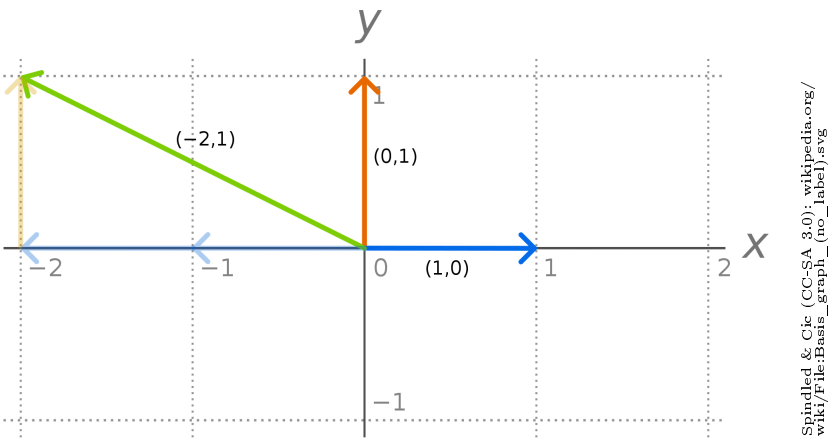


Figure 2.4: Vector (-2,1) in orthonormal basis {(1,0),(0,1)}

in the standard basis (see Fig. 2.4).

This basis yields vectors in terms of the x and y axes, since the basis vector $(1, 0)$ lies on the x-axis, while $(0, 1)$ lies on the y-axis. The list of coefficients for the above linear combination, $(2, 3)$, thus completely defines our vector, *in this basis*. Once we have given our basis, any vector in the space can be defined in terms of this list of coefficients, or weightings for the basis elements, where the number of basis elements is called the "dimensionality" of the vector and the vector space.

In a different basis, the list of coefficients would be different. For instance, using the basis $\{(1, 1), (-1, 1)\}$ would essentially correspond to a 45° rotation of the co-ordinate system:

$$\begin{aligned} (\text{-}\frac{1}{2}, \frac{3}{2}) \quad &\text{in basis} \quad \{(1,1),(-1,1)\} \quad = \\ (-2, 1) \quad &\text{in basis} \quad \{(1,0),(0,1)\} \end{aligned} \tag{2.5}$$

since

$$-\frac{1}{2}(1,1) + \frac{3}{2}(-1,1) = (-2,1) \tag{2.6}$$

Not all alternative bases for this vector space are rotations of the standard basis, or even orthonormal. FIg. 2.5 shows a vector $\mathbf{v}$ that can be decomposed into components vectors of either the orthogonal basis $\{\mathbf{e}_1, \mathbf{e}_2\}$ or the nonorthogonal $\{\mathbf{f_1}, \mathbf{f}_2\}$.

While it is not necessary to express vectors in terms of orthonormal bases, it is customary to do so, and it has many benefits in terms of mathematical elegance. The reader should generally assume that vectors are expressed in terms of an orthonormal basis, unless otherwise explicitly stated.

The basis vectors themselves need not actually be "arrows"—they can be almost any sort of thing, so long as they meet certain requirements that allow vector addition and multiplication by

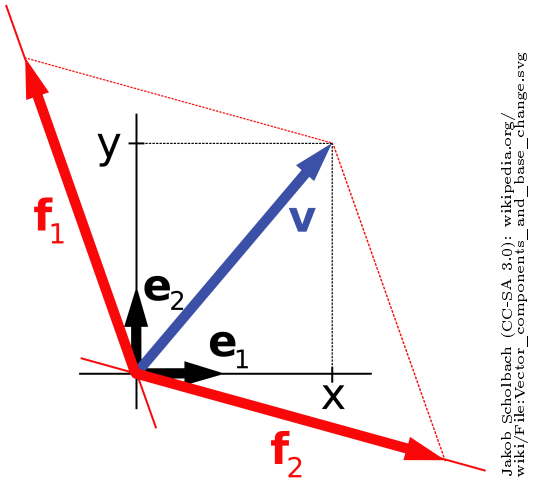


Figure 2.5: A vector represented in two different bases.

scalars in a manner similar to arrows (see [198, p 2]). We will continue to use arrows as a convenient visualization tool, however, even when the discussion is not actually restricted to arrows.

2.2.5 Inner product

In addition to allowing arbitrary (or even infinite) dimensions, Hilbert spaces possess an "inner product" operator that allows for generalizations of the familiar geometrical concepts of angles and distances. For Euclidean arrows expressed in orthonormal bases, the inner product is just the familiar dot product:

$$(x_1, y_1) \cdot (x_2, y_2) \cdot \cdots = x_1 x_2 + y_1 y_2 + \cdots \tag{2.7}$$

In fact, for any Hilbert space, there will always be some orthonormal basis in which the inner product equates to the dot product.

The dot product of $(2, 3)$ and $(5, -4)$ in $\mathbb{R}^2$ (ordered pairs of real numbers) would be

$$\begin{aligned} (2,3) \cdot (5,-4) &= (2 \times 5) + (3 \times -4) \\ &= -2 \end{aligned} \tag{2.8}$$

The dot product can be visualized as measuring the product of the magnitudes of the vectors, weighted by how close the vectors are to being parallel to each other (the cosine function is 0 if the vectors are orthogonal (perpendicular) and 1 if they are parallel):

$$\mathbf{A} \cdot \mathbf{B} = |\mathbf{A}|\,|\mathbf{B}| \cos\theta \tag{2.9}$$

where θ is the angle between the vectors.

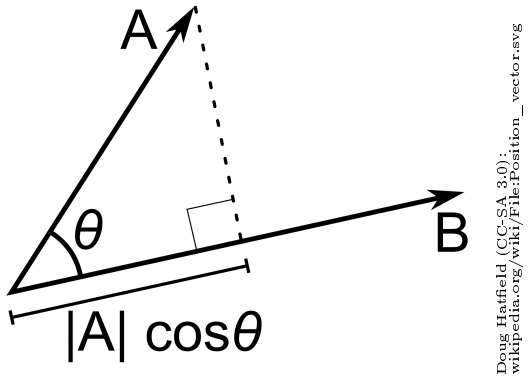


Figure 2.6: Dot product: $\mathbf{A} \cdot \mathbf{B} = |\mathbf{A}|\,|\mathbf{B}| \cos\theta$

The cosine function can be thought of as a measure of how close two vectors are to pointing in the same direction. When the vectors are orthogonal, $\theta = 90°$, $\cos\theta = 0$, and the dot product is zero. In the case of completely parallel vectors, pointing in the same direction, $\theta = 0$, $\cos\theta = 1$, and we are simply dealing with the product of the vectors' magnitudes. If the two vectors are identical, then the inner product is the square of the vector's magnitude:

$$\mathbf{A} \cdot \mathbf{A} = |\mathbf{A}|^2 \tag{2.10}$$

The inner product of $\mathbf{A} \cdot \mathbf{B}$ can also be visualized as the magnitude of $\mathbf{B}$ times the magnitude of the component of $\mathbf{A}$ that is parallel to, or "projects onto" $\mathbf{B}$ (or vice-versa). Imagine this "projection" of $\mathbf{A}$ onto $\mathbf{B}$ as the shadow cast by A onto the ground (on which lies $\mathbf{B}$), with the sun (usually) directly overhead (see Fig. 2.6):

$$\mathbf{A} \dashrightarrow \mathbf{B} = |\mathbf{A}| \cos\theta \tag{2.11}$$

where $\mathbf{A} \dashrightarrow \mathbf{B}$ is the "orthogonal" or "perpendicular projection" of $\mathbf{A}$ onto $\mathbf{B}$.

Generally, projections are perpendicular—and we will generally assume that they are, unless stated otherwise—however, non-perpendicular projections (where the "sun" is not directly overhead) can also be defined.

2.2.6 Complex numbers

In our examples, we will tend to use vectors with real-valued magnitudes, since these are easier to visualize. The kind of numbers our vector space is defined on is called its "field", which we will call $\mathbb{F}$ in general. The Hilbert space $\mathbb{F}^n$ is the n-dimensional Hilbert space over field $\mathbb{F}$. The field $\mathbb{R}$ is the real numbers, and $\mathbb{C}$ is the complex numbers. In general, quantum mechanics uses Hilbert spaces defined over $\mathbb{C}$. However, just as we use arrows to visualize vectors more generally, we will

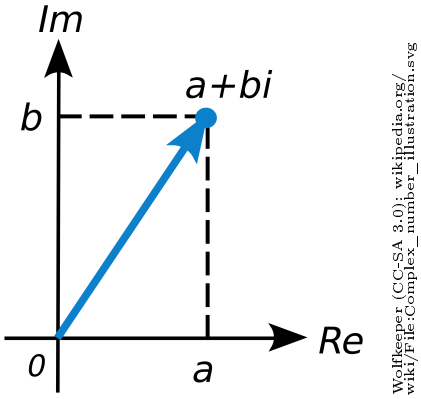


Figure 2.7: Visualization of a complex number

often restrict our geometrical visualization to $\mathbb{R}^3$, simply because $\mathbb{C}^n$ Hilbert spaces are, in general, difficult to visualize. Such simplifications are not without pitfalls, of course, if we forget that, in most real cases, $n > 3$ and $\mathbb{F} = \mathbb{C}$.

Recall that a complex number z has a real component a and an imaginary component b:

$$z = a + bi \tag{2.12}$$

where

$$i = \sqrt{-1} \tag{2.13}$$

A complex number can be visualized as a two-dimensional point on a Cartesian plane, usually with the real (Re) component plotted on the x-axis (the real number line), and the imaginary (Im) component plotted on the y-axis (the imaginary number line) (see Fig. 2.7).

Since a complex number can be decomposed into real and imaginary components, it has a magnitude, r, which is the length of the arrow formed by the head-to-tail sum of the two components. By using the angle this arrow takes from the real number line, we can use the usual trigonometric relationships to express the complex number in trigonometric terms

$$z = \cos\theta + i\sin\theta \tag{2.14}$$

The Pythagorean theorem [78, Bk.1, Pr.47] tells us how to calculate the length of the hypotenuse of the right triangle formed by these two components, which is just the length (magnitude or absolute value) of r:

$$r = |z| = \sqrt{a^2 + b^2} \tag{2.15}$$

This allows our complex number to be expressed in polar co-ordinate form:

$$z = re^{i\theta} = |z|\, e^{i\theta} \tag{2.16}$$

A complex number when used as the amplitude of a wave can serve as a "phase factor". Recall the principles of constructive and destructive interference, as exemplified by the dropping of two pebbles into a pond. Both water waves are, on their own terms, identical, and if we only considered each wave on its own terms, there would be no need to have a phase factor, and we could stick with real numbers to describe the wave. But recall that the two waves can have a phase, relative to each other, resulting in interference. This relative phase can be represented as a complex amplitude. If each wave is considered on its own, we could use its (real) absolute value for the amplitude. But by rotating this amplitude through θ degrees of the complex plane, we can represent that the phase of this wave is θ degrees out of phase with the equivalent real-valued wave. Since the absolute values of the amplitudes of both waves are identical, the two waves represent the exact same physical situation, *on their own*. It is only relative to each other that they have phase, and can interfere with each other when added together.

In quantum mechanics, the wave will be represented by a dynamically changing vector in Hilbert space, with complex amplitudes to represent phase. To rotate the vector in the complex plane, we multiply it by a complex phase factor. Note that it is still possible to have two amplitudes destructively interfere with each other, without using complex numbers. Vectors in $\mathbb{R}^n$ can still have negative amplitudes, so, for instance, a positive amplitude of $+2$ could still destructively interfere with a negative amplitude of -3 to produce a resultant amplitude of -1 (which would be physically equivalent to an amplitude of 1). Thus, even when discussing interference effects, we can still often stick to talking about real-valued amplitudes, since real amplitudes are much easier to visualize. However, one can readily see why, in general, complex amplitudes are needed.

To compute the "complex conjugate", z^*, of a complex number z, one simply negates the imaginary component (see Fig. 2.8):

$$\begin{aligned} z &= a + bi \\ z^* &= a - bi \end{aligned} \tag{2.17}$$

This is equivalent to simply negating the exponent, in the polar form, since the complex conjugate is simply the original complex number rotated from the real number line by the same amount, but in the opposite direction ($-\theta$, clockwise, instead of $+\theta$, counter-clockwise; see Fig. 2.8):

$$z^* = |z|\, e^{-i\theta} \tag{2.18}$$

The polar form makes visualizing complex numbers much easier than does the component form, since it allows us to imagine z as a real magnitude ($|z|$) that is multiplied by a (unmeaningful on its own) phase factor ($e^{i\theta}$). However, keep in mind that this "real value" of the complex number is

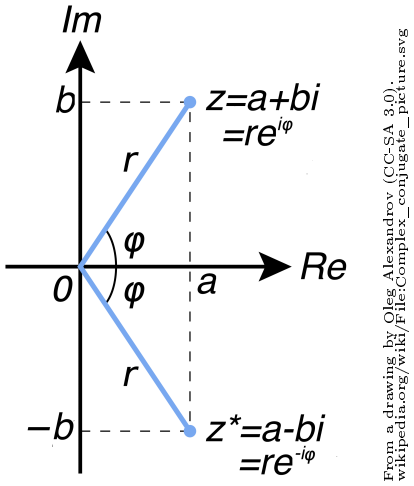


Figure 2.8: Complex conjugation

its *absolute* value, which is not to be confused with its real *component*, a. Because r is a constant, no matter the phase, it is better, intuitively, to think of r as the "real part" of the complex number, rather than a, which depends on the phase. Then, the "imaginary part" is simply a rotation between zero and τ radians (360^{o}), rather than a component, which makes sense if it is to represent the phase of a wave. A factor of i rotates a number through the complex plane, in the same way a factor of -1 reflects a number on the real number line.

2.2.7 Dirac bra-ket notation

From here on, we will use the Dirac notation for vector spaces, which is especially suited to quantum mechanics, and was invented by Paul Dirac for that purpose [72]. It can also be used simply as a general mathematical notation for linear algebra and for Hilbert spaces, so is not necessarily tied to quantum mechanics.

Up to now, we have used a bold font to represent vectors. In Dirac notation, we will put a vector inside a $|\rangle$ or "ket". Such a vector can also simply be called a ket. So instead of writing $\mathbf{A}$, we write

$$|A\rangle$$

We can also put the same vector inside a $\langle|$ or "bra":

$$\langle A|$$

The two vectors $\langle A|$ and $|A\rangle$ represent, each on their own, essentially the same vector. They are only distinguished relative to each other, each being said to be the "dual" of the other, existing in

"dual vector spaces". But those dual spaces are, each on its own terms, formally equivalent to the other.

This is a little like handedness. Your left hand is formally equivalent to your right. Considered, each on its own terms, there is no analytic distinction between the shape of the left and the right hand. However, when you bring your two hands together, there is a difference, and they are now "duals" of each other, but the difference between them lies wholly in their relationship to each other.

Recall that when we are using an orthonormal basis, we can represent a vector in Hilbert space as a list of coefficients (weightings for the basis elements). We can thus treat the n-dimensional vector as if it were an n-dimensional arrow. Thus, so long as we stick to using orthonormal bases, we distinguish vectors in dual spaces firstly by writing the ket as a column vector (or single-column matrix) of scalar coefficients, and the bra as a row vector (or single-row matrix) of scalar coefficients. So the bra is a transpose of the ket—at least, this is all we need to do for reals. So if $\mathbf{A} = (1, 2, 3)$ in $\mathbb{R}^3$, we write the bra and ket like so:

$$|A\rangle = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$
$$\langle A| = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \tag{2.19}$$

Since we can now treat the vectors like arrows, the inner product becomes the dot product. This allows us to define the inner product $\langle A\,|B\,\rangle$ of

$$|A\rangle = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

and

$$|B\rangle = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$$

using matrix multiplication:

$$\begin{aligned} \langle A\,|B\,\rangle &= \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} \\ &= (1)(3) + (2)(2) + (3)(1) \\ &= 10 \end{aligned} \tag{2.20}$$

Note that the inner product in $\mathbb{R}^n$ is commutative. However, more generally, in $\mathbb{C}^n$, the inner product is not commutative, but rather it is "conjugate-symmetric":

$$\langle A|B\rangle = \langle B|A\rangle^* \tag{2.21}$$

This is because the bra is further distinguished from the ket, in $\mathbb{C}^n$, by a phase factor (a reflection about the real axis in the complex plane), so its elements are all the complex conjugates of the ket's elements.

When we transpose a matrix (equivalent to rotating it clockwise 90°, so that its rows become columns and its columns, rows), and then complex conjugate its elements, the result is called the "conjugate-transpose". The conjugate-transpose can be considered as the complex conjugate (for scalars) generalized to vectors and matrices, so we will use the same asterisk, *, to represent it. I will use the term "complex conjugate" as the general term that applies to scalars, vectors and matrices.

The conjugate-transpose of the ket $|\psi\rangle$ is $|\psi\rangle^*$, otherwise known as the bra $\langle\psi|$.

$$\langle\psi| = |\psi\rangle^* \tag{2.22}$$

Any expression can be converted to the dual space expression simply by conjugate-transposing all the vectors, and taking the complex conjugates of all the scalars, while converting kets to bras and bras to kets.

For a quantum wavefunction, converting to the dual space will not change the vector's physical meaning, since the complex conjugate only affects an amplitude by a phase factor, and the difference between a column and row vector is nil unless the two interact in some way (as with the inner product).

In general, we can write a vector $|\psi\rangle$ as a linear combination of basis elements $\{|b_k\rangle\}$ weighted by complex coefficients $\{\psi_k\}$:

$$|\psi\rangle = \sum_k \psi_k |b_k\rangle \tag{2.23}$$

When it is clear which basis we are working in, we can keep things notationally cleaner by writing $|k\rangle$ instead of $|b_k\rangle$.

$$|\psi\rangle = \sum_k \psi_k |k\rangle \tag{2.24}$$

Expressing it as an arrow in an n-dimensional basis:

$$|\psi\rangle = \begin{bmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_n \end{bmatrix} \qquad |k\rangle = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{2.25}$$

where the sole 1 in $|k\rangle$ is in the kth row. So the inner product of the k-th basis element and a ket is the simply the k-th element of the ket:

$$\langle k \,|\, \psi\rangle = \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_n \end{bmatrix} = \psi_k \tag{2.26}$$

So that a basis-element, as a bra, acts simply as an index to the ket it is applied to.

Given another non-basis vector $|\varphi\rangle = \sum_k \varphi_k |k\rangle$, we can write the inner product of it and our first vector as

$$\langle \varphi \,|\, \psi\rangle \quad = \quad \begin{bmatrix} \varphi_1^* & \varphi_2^* & \cdots & \varphi_n^* \end{bmatrix} \begin{bmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_n \end{bmatrix} = \sum_k \varphi_k^* \psi_k \tag{2.27}$$

The "norm" of a vector generalizes the notion of absolute value (or length) from scalars to vectors. We will therefore use the same notation for it (although many people prefer to use double lines, $\| \, \|$, to distinguish it from the absolute value of a scalar):

$$\big||\psi\rangle\big| = \sqrt{\langle\psi|\psi\rangle} \tag{2.28}$$

I will also use "norm" as the general term that applies to both scalars and vectors.

In addition to choosing orthonormal bases, it is customary in quantum mechanics to "normalize" vectors, by dividing vectors by their norms, and it is usually assumed that this has been done unless specified otherwise. Normalizing vector $|\psi\rangle$ yields

$$\frac{|\psi\rangle}{\big||\psi\rangle\big|} \tag{2.29}$$

The inner product of a normalized vector $|\psi\rangle$ with itself will always be unity:

$$\langle\psi|\psi\rangle = 1 \tag{2.30}$$

One can clearly see that this is the case for an orthonormal basis element, consisting of all 0's and one 1 (and it is not difficult to generalize from that fact to the more general case).

2.2.8 Operators

Definition 2.1. An "operator" takes a ket (or bra) and transforms it in some way, producing another ket (or bra).

An operator might scale, rotate, or do any manner of transformation to the ket (although not all transformations are allowed in quantum mechanics). In Dirac notation, operators are normally capitalized and placed to the *left* of the ket (or the *right* of the bra) that they act on. So operator O acting on ket $|\psi\rangle$ to produce $|\psi'\rangle$ can be represented with

$$|\psi'\rangle = O\,|\psi\rangle \tag{2.31}$$

Note that the scalar in scalar multiplication can be considered a kind of operator that simply scales the ket. Although, in general, an operator acts on a ket to produce another ket, we can view a bra (or ket) as a kind of operator that produces, instead of another ket (or bra), a scalar via the inner product.

If $\hat{R}$ is the "rotate 90° clockwise" transformation, we would indicate a rotation on ket $|\psi\rangle$ with

$$|\psi'\rangle = \hat{R}\,|\psi\rangle \tag{2.32}$$

A trivial (but still a very useful) operator is the identity operator, $\hat{I}$, which leaves a vector unchanged:

$$|\psi\rangle = \hat{I}\,|\psi\rangle \tag{2.33}$$

Note that the scalar 1 is essentially the same thing as $\hat{I}$, when considered as an operator. Operators can be chained, with the first to be applied to the ket situated furthest to the right. So we could write

$$|\psi'\rangle = \hat{I}\hat{R}\,|\psi\rangle \tag{2.34}$$

to indicate the rotation of $|\psi\rangle$, followed by the application of identity. Note that operators do not necessarily commute: the order in which they are applied can make a difference. Of course, in the above example, since the second operation is identity, it actually does *not* matter what order the operators are applied (since $\hat{I}$ commutes with $\hat{R}$, as it does with everything else). But in general it *does* matter.

Definition 2.2. A "distributive operator" $\tilde{L}$ is defined as any operator that satisfies

$$\begin{aligned} \tilde{L}\left(|\psi\rangle + |\varphi\rangle\right) &= \tilde{L}\,|\psi\rangle + \tilde{L}\,|\varphi\rangle \\ \left(\langle\psi| + \langle\varphi|\right)\tilde{L} &= \langle\psi|\,\tilde{L} + \langle\varphi|\,\tilde{L} \end{aligned} \tag{2.35}$$

In other words, a distributive operator can be applied to a vector either as a whole, or piecewise to its components, and the result will be the same either way. I indicate that an operator is distributive by placing a tilde $\sim$ above the letter. All operators in quantum mechanics are distributive.

There are two kinds of distributive operators relevant to quantum mechanics: linear and antilinear.

Definition 2.3. A "linear operator" $\hat{L}$ is defined as a distributive operator that satisfies, for any scalar a,

$$\hat{L}a = a\hat{L} \tag{2.36}$$

Definition 2.4. An "antilinear" or "conjugate-linear" operator $\check{L}$ is defined as a distributive operator that satisfies

$$\check{L}a = a^*\check{L} \tag{2.37}$$

I will write linear operators with a "hat" or caret symbol, and use an upside-down caret for antilinear operators, following [181]. Both types of operators could be written with the tilde instead—since they are both also distributive—but the tilde would normally only be used if we did not wish to commit to whether the operator was linear or antilinear, so it will not be used much in practice.

A linear operator is essentially an operator that acts equivalently on wholes or on parts (it is distributive), and for which the scalars in scalar multiplication act as universally-commutative linear operators (they commute with everything). Anti-linear operators act similarly, except that scaling is no longer universally-commutative, as we must use the complex conjugate of the scalar when reversing the order in which the scalar and the operator are applied. Operators in quantum mechanics are linear, at least as applied to empirical phenomena (sometimes antilinear operators may be used for analytical purposes, as in the case of the time-reversal operator).

Definition 2.5. When an operator $\hat{O}$ has the exact same effect on a ket $|\psi\rangle$ as multiplying by some scalar o,

$$\hat{O}\,|\psi\rangle = o\,|\psi\rangle \tag{2.38}$$

we call $|\psi\rangle$ an "eigenket" or "eigenvector" of $\hat{O}$, and o is its "eigenvalue" corresponding to this eigenket.

Definition 2.6. If each eigenket has a unique eigenvalue, then the eigenvalues are "nondegenerate" and there are as many eigenvalues as eigenkets. If some of the eigenvalues are the same, however, then we call the duplicated eigenvalues "degenerate". The "multiplicity" of a eigenvalue is the number of different eigenkets it has.

When representing kets as matrices, a linear operator can be defined as the "outer product" (matrix multiplication) of a ket times a bra (simply the reverse order of the inner product). Thus, the following is an operator:

$$|\varphi\rangle \langle\phi| \tag{2.39}$$

and produces an $n \times n$ matrix. This all works out, because, just as a row vector times a column vector yields a scalar, a column vector times a row vector yields a square matrix. In addition, a square matrix times a column vector yields another column vector (a transformed ket). All this follows from the distinction between column and row vectors, treating them as matrices, and subsuming both the inner and outer products as instances of matrix multiplication.

We can take the conjugate-transpose of an operator matrix, just as we can for the bras and kets. We can likewise continue to use the asterisk notation, $\hat{O}^*$ (although many people prefer to use the dagger symbol, $\hat{O}^\dagger$, to distinguish it from the conjugate transpose of a vector). Converting expressions to their dual form follows the same rules as before, but with the addition of performing conjugate-transpose on all the operators.

Expanding the operator $\hat{O} = |\varphi\rangle \langle\phi|$, in matrix form, gives us

$$\begin{aligned}
\hat{O} = |\varphi\rangle \langle\phi| &= \begin{bmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_n \end{bmatrix} \begin{bmatrix} \phi_1^* & \phi_2^* & \cdots & \phi_n^* \end{bmatrix} \\
&= \begin{bmatrix} \varphi_1\phi_1^* & \varphi_1\phi_2^* & \cdots & \varphi_1\phi_n^* \\ \varphi_2\phi_1^* & \varphi_2\phi_2^* & \cdots & \varphi_2\phi_n^* \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_n\phi_1^* & \varphi_n\phi_2^* & \cdots & \varphi_n\phi_n^* \end{bmatrix} \\
&= \begin{bmatrix} \ddots & \vdots & \iddots \\ \cdots & \varphi_r\phi_c^* & \cdots \\ \iddots & \vdots & \ddots \end{bmatrix} = \begin{bmatrix} \ddots & \vdots & \iddots \\ \cdots & \left\langle r|\hat{O}|c \right\rangle & \cdots \\ \iddots & \vdots & \ddots \end{bmatrix}
\end{aligned} \tag{2.40}$$

The operator can be thought of as making the following request of whatever ket it is applied to:

$|\varphi\rangle\langle\phi|$ = *"Scale me with our overlap."* E.g.:
$|\varphi\rangle\langle\phi|\psi\rangle$ = *"Hello* $|\psi\rangle$*, scale my ket* $|\varphi\rangle$ *by the amount of your overlap with my bra* $\langle\phi|$*."*

Recall that a basis element, as a bra in an inner product, acts as an index operator. So:

$|\varphi\rangle\langle k|$ = *"Scale me with your k-th component."* E.g.:
$|\varphi\rangle\langle k|\psi\rangle$ = *Hello* $|\psi\rangle$*, scale my ket* $|\varphi\rangle$ *by your component for the k-th basis element.*

The projection of one ket onto another is expressed very naturally in Dirac notation, as the outer product of the ket/column form of a vector with its bra/row form. So the orthogonal projection operator $\dashrightarrow \varphi$ becomes, in Dirac notation, $|\varphi\rangle\langle\varphi|$. Projecting (casting the shadow of) vector $|\psi\rangle$ onto vector $|\varphi\rangle$ gives us

$$|\psi\rangle \dashrightarrow |\varphi\rangle = |\varphi\rangle\langle\varphi|\,\psi\rangle \tag{2.41}$$

We will call the orthogonal projection operator $|\varphi\rangle\langle\varphi|$ the "projector" for $|\varphi\rangle$. We will sometimes use a short-hand notation (following [99]) for a projector, so that

$$[\varphi] \quad = \quad |\varphi\rangle\langle\varphi| \tag{2.42}$$

A projector (or any projection operator, for that matter) is necessarily "idempotent", meaning that it can be applied twice in succession (or any number of times) and will always produce the same result (since the original vector has already been projected onto a certain subspace, further projection onto that same subspace will leave the result untouched).

$$\begin{aligned} [\varphi]^2 &= [\varphi] \\ [\varphi]\,[\varphi] &= [\varphi] \\ \cdots[\varphi]\,[\varphi]\,[\varphi] &= [\varphi] \end{aligned} \tag{2.43}$$

For a projector, the state (ket) that is scaled is the same as the state (bra) projected onto, so the projector's request becomes:

$|\varphi\rangle\langle\varphi|$ = *"Give me my component."* E.g.:
$|\varphi\rangle\langle\varphi|\psi\rangle$ = *"Hello* $|\psi\rangle$*, give me your component lying in my direction."*

The component that the projector $[\varphi]$ gets back from its request is its state $|\varphi\rangle$ multiplied by a scale factor ψ_φ (so that $[\varphi]\,|\psi\rangle = \psi_\varphi\,|\varphi\rangle$). This scale factor must not be confused with the eigenvalue, call it e_ψ, of an eigenket $|\psi\rangle$ of the projector $[\varphi]$, which is also a scale factor (so that $[\varphi]\,|\psi\rangle = e_\psi\,|\psi\rangle$.). An eigenvalue scales the original ket it was applied to, not the unit ket of the subspace projected on.

Clearly, not all kets will be eigenkets of a given projector. For $|\psi\rangle$ to be an eigenket of $[\varphi]$, the projector $[\varphi]$ must both *(1)* project onto the subspace of $|\psi\rangle$, and *(2)* merely scale $|\psi\rangle$. The only way to project, while at the same time doing nothing more than scaling, is either to:

1. Leave the ket alone (scale by 1 while projecting onto the same ket's subspace), or
2. Zero the ket entirely (scale by 0 while projecting onto an orthogonal subspace).

Hence, the eigenvalues for a projector are *always* either 0 or 1. So when applied to eigenket $|\psi\rangle$, the projector's request becomes the simple question:

$|\varphi\rangle\langle\varphi|\psi\rangle =$ *"Hello my eigenket, are you me?"*

And since $|\psi\rangle$ is an eigenket, the answer will be "yes" or "no". For the projector of a basis element, $[k]$, the eigenkets will be the orthogonal basis elements. So when applied to an eigenket, the question becomes

$|k\rangle\langle k|i\rangle = [k]\,i\rangle = |k\rangle$ or $0 =$ *"Hello eigenket, are you the k-th basis element?"*

When applied to a non-eigenket $|\psi\rangle$, the projector will *not* scale by 0 or 1, but since this state (like any state in the Hilbert space) can be analyzed as a linear combination of basis elements, the projector can be interpreted as asking each of the components in turn the same yes-no question. However, since the basis is orthogonal, only one of the answers will be "yes".

$|k\rangle\langle k| = [k] =$ *"Do you contain me?"*
$|k\rangle\langle k|\psi\rangle = [k]\,\psi\rangle = [k]\,(\sum_i \psi_i\,|i\rangle) = \psi_k\,|k\rangle =$ *"Hello components of $|\psi\rangle$, are you me?"*

Thus, in general, we can view the projector either as *(1)* requesting its own component back, or *(2)* asking whether its component is present. The former view considers the "result" of the projection to be the *final projected state*, while the latter sees the *eigenvalue* as the "result" of the projection (in neither case is the basis element's scale factor, ψ_k, considered to be the result). Both are valid ways of talking about projection; the main thing is not to confuse them, since they both involve (completely different) scale factors.[9]

If we consider $[k]$ to define (test for) a certain property, then we can reword the question as

$|k\rangle\langle k| = [k] =$ *"Do you have property k?"* E.g.:
$|\text{blue}\rangle\langle\text{blue}| = [\text{blue}] =$ *"Are you blue?"*

If we project a ket onto the subspace defined by the kth basis element, $|k\rangle$, we get the kth element of the vector in that basis, which can be considered as a weighting on the k-th basis element. If we

[9] In quantum mechanics, the standard model of measurement is "projective" measurement, which takes a measurement to be the application of a projector, and the observed measurement result as the resulting *eigenvalue*, so that it is view #2 that corresponds to a measurement "outcome", such as a pointer reading, while view #1 corresponds to the state the system is left in after the measurement.

sum up all n of these components, for the entire n-dimensional basis, we will have re-constructed the original vector:

$$\sum_k [k]\,\psi\rangle = \sum_k |k\rangle\,\langle k|\,\psi\rangle = \sum_k \langle k|\,\psi\rangle\,|k\rangle = \sum_k \psi_k\,|k\rangle = |\psi\rangle \tag{2.44}$$

Thus, the n vectors in an orthonormal basis for an n-dimensional Hilbert space yield projectors which, when summed, are identical to the n-dimensional identity operator for the space:

$$\sum_k [k] = \hat{I} \tag{2.45}$$

So far, our projectors have been 1-dimensional, since they project onto a 1-D subspace, or ray, of the Hilbert space (one of the basis elements). Higher-dimensional subspaces of the Hilbert space will also yield projectors. Any subset of m of these basis projectors can be considered the identity operator for the m-dimensional subspace defined by taking those basis elements as an orthonormal basis. By summing two 1-D basis projectors, we can create a 2-D projector that projects onto a 2-D subspace of the Hilbert space, which we can abbreviate:

$$[i,j] = [i] + [j] \tag{2.46}$$

This corresponds to combining questions together, which we can do, so long as each question asks about something orthogonal, so we can ask the questions simultaneously in one combined projector. For instance

$|i\rangle\,\langle i| + |j\rangle\,\langle j| = [i,j]$= *"Do you have properties i and j?"* E.g.:
$|\text{blue}\rangle\,\langle\text{blue}| + |\text{red}\rangle\,\langle\text{red}| = [\text{blue,red}]$= *"Are you blue and red?"*

Note that there is something funny about interpreting a projector as a question. Asking if the system is blue will leave the system as either blue or zero, even if it contains red. Asking if the same system is blue *and* red will leave it as blue and red (eliminating any other orthogonal colors). This is because a projector doesn't just *answer* the question, it also throws out all the orthogonal alternatives in the process of giving its answer.

Definition 2.7. The projector $\left[\bar{\psi}\right]$ that collects together all the members of a basis *except* the members of another projector $[\psi]$, is the "complement" of that other projector (and we will likewise say that $\left|\bar{\psi}\right\rangle$ is the complement of $|\psi\rangle$). For example, the complement of $[k]$ is $[1,\ldots,k-1,k+1,\ldots,n]$ in the n-dimensional basis $\{|1\rangle,\ldots,|n\rangle\}$.

Theorem 2.8. *The complement* $\left[\bar{\psi}\right]$ *of a projector* $[\psi]$ *is basis-independent.*

Proof. Clearly, $|\psi\rangle$ is the same vector no matter what basis you express it in. If $[\psi] = [1, \ldots, k, \ldots, n_\psi]$ in a basis, and $\left[\bar{\psi}\right] = \left[\bar{1}, \ldots, \bar{k}, \ldots, \bar{n}_\psi\right]$, then clearly $\left[1, \ldots, k, \ldots, n_\psi, \bar{1}, \ldots, \bar{k}, \ldots, \bar{n}_\psi\right] = \hat{I}$, and $\left|\bar{\psi}\right\rangle$ will also be the same vector no matter what basis we choose, since it must, together with $|\psi\rangle$, span the entire space. □

Now imagine a vector in our n-dimensional space that happens to lie on the plane of the 2-D subspace for $[i, j]$:

$$c_i |i\rangle + c_j |j\rangle \tag{2.47}$$

Apply the 2-D projector $[i, j]$, recalling that since our basis is orthonormal, $\langle k|\ k\rangle = 1$ and $\langle i|\ j\rangle = 0$ when $i \neq j$:

$$\begin{aligned} & ([i, j]) (c_i |i\rangle + c_j |j\rangle) \\ = \ & c_i |i\rangle \langle i|\ i\rangle + c_i |j\rangle \langle j|\ j\rangle + c_j |i\rangle \langle i|\ j\rangle + c_j |j\rangle \langle j|\ j\rangle \\ = \ & c_i |i\rangle + c_j |j\rangle \end{aligned} \tag{2.48}$$

which is just the original vector.

If we have a vector lying wholly in a 3-D subspace of the Hilbert space:

$$c_i |i\rangle + c_j |j\rangle + c_k |k\rangle \tag{2.49}$$

and we project it onto the same 2-D subspace:

$$\begin{aligned} & ([i, j]) (c_i |i\rangle + c_j |j\rangle + c_k |k\rangle) \\ = \ & c_i |i\rangle \langle i|\ i\rangle + c_j |i\rangle \langle i|\ j\rangle + c_k |i\rangle \langle i|\ k\rangle \\ + \ & c_i |j\rangle \langle j|\ i\rangle + c_j |j\rangle \langle j|\ j\rangle + c_k |j\rangle \langle j|\ k\rangle \\ = \ & c_i |i\rangle + c_j |j\rangle \end{aligned} \tag{2.50}$$

which is just the component of the original vector lying in our subspace.

Note that when we construct higher dimensional projectors, we sum 1-D projectors:

$$[i, j, \ldots] = [i] + [j] + \cdots \tag{2.51}$$

We do *not* chain them:

$$[i, j, \ldots] \neq [i]\,[j] \cdots \tag{2.52}$$

Chaining, rather than summing, the same two basis elements as above, on the same vector lying on

a 3-D subspace, yields the (n-dimensional) null or zero vector:

$$\begin{aligned} & ([i,j])\left(c_i\left|b_i\right\rangle + c_j\left|b_j\right\rangle + c_k\left|b_k\right\rangle\right) \\ = \; & \left(\left|i\right\rangle\left\langle i\right|\, j\right\rangle\left\langle j\right|)\left(c_i\left|i\right\rangle + c_j\left|j\right\rangle + c_k\left|k\right\rangle\right) \\ = \; & 0\left(c_i\left|i\right\rangle + c_j\left|j\right\rangle\right) \\ = \; & \left|zero\right\rangle = 0 \end{aligned} \tag{2.53}$$

This is because chaining the projectors means repeatedly applying them one after the other. If I consecutively project onto two orthogonal 1-D subspaces, I am left with nothing (but only because they are orthogonal to each other).

Each possible way of partitioning the basis set into mutually exclusive and exhaustive subsets yields a particular "refinement" or "decomposition of identity" for that basis, yielding an "orthogonal collection" of projectors, that together sum to identity for the whole space. The projectors in such a collection can be of differing dimensionality. For instance, for our 3-D Hilbert space, we could create the decomposition $\{[i,j],k\}$, consisting of one 2-D and one 1-D projector, which together cover the 3-D space.

Note that the $2-$D projector above is only "two-dimensional" in this particular basis. There is another basis for the Hilbert space in which the unit norm 2-D vector defining this projector is a 1-D vector, and a single element of *that* basis set.

2.2.9 Hermitian Operators

An operator $\hat{A}$ is "Hermitian" if it is complex conjugate-symmetric:

$$\hat{O} = \hat{O}^* \tag{2.54}$$

This is the complex generalization of a symmetric matrix of reals (*i.e.*, where the matrix is simply equal to its transpose). The following operator matrix, for example, is Hermitian:

$$\begin{bmatrix} 1 & 4+i & 5+i \\ 4-i & 2 & 6+i \\ 5-i & 6-i & 3 \end{bmatrix} \tag{2.55}$$

Note that the main diagonal of an Hermitian matrix is always real. This is because these numbers map to themselves in the conjugate transpose, and so they must be equal to their own complex conjugates (which is the case only for reals). The off-diagonals (in general) have imaginary components (in other words, additional phase factors). An "Hermitian" projection operator is simply an orthogonal one, or a projector.

It is not difficult to show (see [198, p 35]) that the eigenvalues of an Hermitian operator are necessarily real. Keep in mind that the actual form the operator matrix takes will depend on the chosen basis. It can be shown (see [198, p 36]) that, for any Hermitian operator, there exists an orthonormal basis that "diagonalizes" the operator, meaning that the operator matrix will be diagonal (with real-valued diagonal elements and all zero off-diagonals) in that basis:

$$\left\langle i|\hat{A}|j\right\rangle = \begin{cases} 0 & \text{if } i \neq j \\ \text{a real number} & \text{if } i = j \end{cases} \tag{2.56}$$

The basis elements $\{|k\rangle\}$ will be exactly the eigenkets of the operator, and the operator can be written as:

$$\hat{A} = \sum_k a_k \left[k\right] \tag{2.57}$$

where the $\{a_k\}$ are the eigenvalues of $\hat{A}$ corresponding to the basis elements or operator eigenkets. Such a basis is a "diagonal basis" of the Hermitian operator. It is always possible (although not always desirable) to generate a "unique diagonal basis" from an operator, which yields unique (non-degenerate) eigenvalues. While it is still possible to diagonalize the operator degenerately, one can simply collect the projectors with equal eigenvalues together, so if $[i]$ and $[j]$ both have the same eigenvalues, we just replace them with $[i, j]$.

Definition 2.9. A "positive semidefinite operator" is an Hermitian operator $\hat{A} = \sum_k a_k \left[k\right]$, for which all its eigenvalues are non-negative,

$$\forall k : a_k \geq 0 \tag{2.58}$$

$$\left\langle k|\hat{A}|k\right\rangle \geq 0 \tag{2.59}$$

from which it is easy to show more generally that

$$\left\langle \psi|\hat{A}|\psi\right\rangle \geq 0 \tag{2.60}$$

Definition 2.10. The "trace" of an operator, Tr(), is linear and independent of the basis. It is defined as the sum of its diagonal terms:

$$\mathrm{Tr}(\hat{O}) = \sum_k \left\langle k\right| \hat{O} \left|k\right\rangle \tag{2.61}$$

Definition 2.11. A "density matrix" (or, more correctly, "density operator") is a non-zero positive operator with unit trace (*i.e.*, normalized so that its diagonal sums to 1), so that it can be expressed in the form

$$\frac{\hat{A}}{\mathrm{Tr}(\hat{A})} \tag{2.62}$$

2.2.10 Operator commutativity

Since operators do not in general commute, $\hat{O}\hat{R}$ is not necessarily the same operator as $\hat{R}\hat{O}$. When two operators *do* commute, we say they are "compatible" with each other. Compatibility can be considered a matter of degree—rather than an all-or-nothing property—by defining it as a measure of the "degree of commutativity" (where strict commutativity is still considered all-or-nothing).

Definition 2.12. Define the "degree of commutativity" or "compatibility" $\left[\hat{O};\hat{R}\right]$ of operators $\hat{O}$ and $\hat{R}$ such that:

$$\left[\hat{O};\hat{R}\right] = \hat{O}\hat{R} - \hat{R}\hat{O} \tag{2.63}$$

Strict commutativity, then, is indicated by the condition

$$\left[\hat{O};\hat{R}\right] = 0 \tag{2.64}$$

while any degree of incompatibility implies that

$$\left[\hat{O};\hat{R}\right] \neq 0 \tag{2.65}$$

However, we can still say, when a non-zero compatibility measure is near-zero,

$$\left[\hat{O};\hat{R}\right] \approx 0 \tag{2.66}$$

that the two operators are "highly compatible" or "approximately commutative".

Note that projectors from the same basis always commute, since they are mutually orthogonal. For example,

$$\begin{aligned} [[i]\,;[j]] &= 0 \\ [[i,j]\,;[k]] &= 0 \\ [[i,j]\,;[i]] &\neq 0 \end{aligned} \tag{2.67}$$

When two Hermitian operators, $\hat{A}$ and $\hat{B}$ commute with each other, this means that they can share the same diagonal basis, so that both can be expressed as a linear combination of projectors in the same basis, weighted by their eigenvalues:

$$\begin{aligned} \hat{A} &= \sum_k a_k\,[k] \\ \hat{B} &= \sum_k b_k\,[k] \end{aligned} \tag{2.68}$$

If we express each in terms of its *unique* diagonal basis,

$$\begin{aligned} \hat{A} &= \sum_k a_k\,[k] \\ \hat{B} &= \sum_{k'} b_{k'}\,[k'] \end{aligned} \tag{2.69}$$

then the bases may not be the same, since $[i]$ and $[j]$ in one basis might, for instance, be combined into $[k'] = [i, j]$ in the other. However, the projectors in the two bases will clearly still commute with each other:

$$\begin{aligned} [i]\,[j'] &= [j']\,[i] \\ [[i]\,;[j']] &= 0 \end{aligned} \tag{2.70}$$

2.2.11 Unitary operators

Definition 2.13. A "symmetry operator" is a distributive operator $\tilde{U}$ that satisfies

$$\tilde{U}^*\tilde{U} = \tilde{U}\tilde{U}^* = \hat{I} \tag{2.71}$$

In other words, the operator commutes (is compatible with) its own conjugate transpose, both operators being each others' inverse:

$$\left[\tilde{U}^*;\tilde{U}\right] = 0 \tag{2.72}$$

An important consequence of this property is that symmetry operators preserve structure, so they are sometimes called "information-preserving" operators.

There are two (and *only* two [232]) kinds of symmetry operators on Hilbert spaces: unitary and antiunitary (see also [12]).

Definition 2.14. A "unitary" operator is defined as a *linear* symmetry operator.

Definition 2.15. An "antiunitary" or "conjugate-unitary" operator is defined as an *antilinear* symmetry operator.

A unitary operator $\hat{U}$ preserves structure by conserving inner products:

$$\left\langle \hat{U}\varphi \,\middle|\, \hat{U}\psi \right\rangle = \langle \varphi \,|\psi\rangle \tag{2.73}$$

while an antiunitary operator $\check{U}$ does so by conjugating them:

$$\left\langle \check{U}\varphi \,\middle|\, \check{U}\psi \right\rangle = \langle \varphi \,|\psi\rangle^* \tag{2.74}$$

Either way, structure or information is preserved. A simple example of a structure-preserving transformation is a rotation operator, which clearly leaves the structure of what it is rotating alone, and indeed symmetry transformations can be considered as a generalization of rotation transformations.

In quantum mechanics, forward-in-time evolution of the wavefunction is unitary, while time-reversed evolution is antiunitary.[10] Both forward and backward temporal transformations preserve

[10] That time-reversal equates to antiunitary evolution makes intuitive sense, since rotations through the complex plane are not properties of systems but represent phase relationships between subsystems—for wavefunction evolution this means *timing* relationships—and to reverse the direction (but not speed) of rotation through the complex plane of one subsystem compared to another would be precisely to reverse its timing.

all the information in the wavefunction, since both are symmetry transformations. Of course, since we perceive the evolution of our world as forward-in-time, it is standard to do quantum mechanics exclusively with unitary transformations.

In quantum mechanics, since the Born rule is a measure on amplitudes (an "amplitude-counting" measure), and only symmetry operators preserve amplitude counts (norm-squared amplitudes), it follows that only symmetry operators will conserve probabilities under the Born rule (this does not mean that unitary evolution implies the Born rule, only that the Born rule implies unitary, or else antiunitary, evolution).

2.2.12 Tensor product operators

Not only does every possible state of S yield a projector (which in quantum mechanics, will become an observable), but any possible state of any *subsystem* of S (in a tensor factor space of the Hilbert space) can also (effectively) be used as a projector (and hence, in quantum mechanics, an observable). Say that our system state $|\psi\rangle$ is factorable into two subsystems S_A, currently in state $|\psi_\alpha\rangle$, and S_B, currently in state $|\psi_\beta\rangle$, each represented in tensor factor spaces $\mathcal{H}_\mathcal{A}$ and $\mathcal{H}_\mathcal{B}$ of $\mathcal{H}_\mathcal{S}$, respectively, so that $\mathcal{H}_\mathcal{S} = \mathcal{H}_\mathcal{A} \otimes \mathcal{H}_\mathcal{B}$.

Any pair of operators $\hat{A}$ and $\hat{B}$ on the two factor spaces $\mathcal{H}_\mathcal{A}$ and $\mathcal{H}_\mathcal{B}$ can be formed into a tensor product operator, $\hat{A} \otimes \hat{B}$, which follows the expected rule:

$$\left(\hat{A} \otimes \hat{B}\right)(|\psi_\alpha\rangle \otimes |\psi_\beta\rangle) = \hat{A}\,|\psi_\alpha\rangle \otimes \hat{B}\,|\psi_\beta\rangle \tag{2.75}$$

However, if we want to apply the tensor factor operator $\hat{A}$ to the entire system, but leave the other subsystem S_B untouched, we simply create a tensor product operator using $\hat{I}_B$ the identity operator for S_B:

$$\begin{aligned}\left(\hat{A} \otimes \hat{I}_B\right)(|\psi_\alpha\rangle \otimes |\psi_\beta\rangle) &= \hat{A}\,|\psi_\alpha\rangle \otimes \hat{I}_B\,|\psi_\beta\rangle \\ &= \hat{A}\,|\psi_\alpha\rangle \otimes |\psi_\beta\rangle\end{aligned} \tag{2.76}$$

When it creates no confusion, we will act as if we can simply apply the factor operator $\hat{A}$ to the whole system by writing

$$\hat{A}\,|\psi\rangle = \hat{A}\,(|\psi_\alpha\rangle \otimes |\psi_\beta\rangle) = \hat{A}\,|\psi_\alpha\rangle \otimes |\psi_\beta\rangle \tag{2.77}$$

However, it should always be kept in mind that this is a short-hand for writing out the entire product operator. Confusion can result if we do not keep this distinction in mind, especially if $\hat{A}$ and $\hat{B}$ are identical operators, differing only in the space they are defined on.

If a tensor product operator is applied to an entangled state, we get the expected (entangled) result:

$$\left(\hat{A}\otimes\hat{B}\right)\left(c_1\left|\psi_\alpha\right\rangle\otimes\left|\psi_\beta\right\rangle+c_2\left|\varphi_\alpha\right\rangle\otimes\left|\varphi_\beta\right\rangle\right)=c_1\hat{A}\left|\psi_\alpha\right\rangle\otimes\hat{B}\left|\psi_\beta\right\rangle+c_2\hat{A}\left|\psi_\alpha\right\rangle\otimes\hat{B}\left|\psi_\beta\right\rangle \tag{2.78}$$

In general, then, for the application of any tensor product operator we have:

$$\left(\hat{A}\otimes\hat{B}\right)\sum_k\left(c_k\left|\psi_k^A\right\rangle\otimes\left|\psi_k^B\right\rangle\right)=\sum_k\left(c_k\hat{A}\left|\psi_k^A\right\rangle\otimes\hat{B}\left|\psi_k^B\right\rangle\right) \tag{2.79}$$

where the superscripts are used to represent the Hilbert space the ket is represented in.

2.2.13 Probabilities and expectation values

For an operator $\hat{O}$, any given ket $|\psi\rangle$ can be decomposed into normalized eigenkets of $\hat{O}$:

$$\left|\varphi\right\rangle=\sum_k c_k\left|k\right\rangle=\sum_k\left|k\right\rangle\left\langle k\right|\left.\varphi\right\rangle \tag{2.80}$$

When we operate on a noneigen ket:

$$\hat{O}\left|\psi\right\rangle=\left|\varphi\right\rangle \tag{2.81}$$

we get

$$\left|\varphi\right\rangle=\sum_k\left|k\right\rangle\left\langle k\right|\left.\varphi\right\rangle=\sum_k\left\langle k\right|\left.\varphi\right\rangle\left|k\right\rangle=\sum_k\varphi_k\left|k\right\rangle \tag{2.82}$$

The inner product, $\varphi_k=\langle k|\left.\varphi\right\rangle$, gives the weighting for the kth eigenket in the composition of state $|\varphi\rangle$ as a linear combination of eigenkets. As such, it should be able to serve as the basis for a calculation of a probability measure for $|k\rangle$, given state $|\varphi\rangle$. A probability measure needs to be normalized so as to sum to unity. We already know that, for normalized kets

$$\begin{aligned}\left\langle\varphi|\varphi\right\rangle&=1\\&=\sum_k\left\langle\varphi\right|\left.k\right\rangle\left\langle k\right|\left.\varphi\right\rangle\end{aligned} \tag{2.83}$$

Note that, even though the inner product is not commutative, $\langle A|B\rangle\langle B|A\rangle=\langle A|B\rangle\langle A|B\rangle$. Therefore,

$$\sum_k\left\langle\varphi\right|\left.k\right\rangle\left\langle k\right|\left.\varphi\right\rangle=\sum_k\left|\left\langle k\right|\left.\varphi\right\rangle\right|^2=1 \tag{2.84}$$

Thus, it appears that we have a basis for a probability measure:

$$p(k|\varphi)=\left|\left\langle k\right|\left.\varphi\right\rangle\right|^2=\left\langle\varphi\right|\left.k\right\rangle\left\langle k\right|\left.\varphi\right\rangle \tag{2.85}$$

If our ket $|\varphi\rangle$ is not normalized (although we will generally assume normalization), then $\langle\varphi|\varphi\rangle$ is not necessarily unity, but then $\frac{1}{\langle\varphi|\varphi\rangle}$ will serve as a normalization constant, and our probability measure

will be

$$p(k|\varphi) = \frac{|\langle k|\ \varphi\rangle|^2}{\langle\varphi|\varphi\rangle} \tag{2.86}$$

So long as this norm-squared measure obeys the other rules of probabilities, it should be a well-behaved probability measure (and we will see in a later chapter that it does). This is, of course, essentially the Born rule, introduced here as a rather obvious probability measure on Hilbert spaces (although this in no way *proves* the rule for any given empirical application, such as quantum mechanics, so the above will not serve in itself as a response to the Born rule objectors).

When we actually apply $\hat{O}$ to one of its eigenkets, this is just a special case of the above:

$$\hat{O}\,|k\rangle = o_k\,|k\rangle \tag{2.87}$$

and we get

$$\begin{aligned} |k\rangle &= \sum_k |k\rangle\,\langle k|\ k\rangle \\ &= \hat{I}\,|k\rangle \\ &= |k\rangle \end{aligned} \tag{2.88}$$

and

$$p(k|k) = 1 \tag{2.89}$$

$$p(i \neq k|k) = 0 \tag{2.90}$$

Definition 2.16. In general, we can say that the "expectation value" $\left\langle \hat{O} \right\rangle$ of operator $\hat{O}$ (the average result over repeated trials) will be the sum of the operator's eigenvalues weighted by their probabilities:

$$\begin{aligned} \left\langle \hat{O} \right\rangle &= \sum_k o_k p(k|\varphi) \\ &= \sum_k o_k\,|\langle k|\ \varphi\rangle|^2 \\ &= \sum_k o_k\,\langle k|\ \varphi\rangle\,\langle\varphi|\ k\rangle \\ &= \sum_k o_k\,\langle\varphi|\ k\rangle\,\langle k|\ \varphi\rangle \\ &= \sum_k \langle\varphi|\,\hat{O}\,|k\rangle\,\langle k|\ \varphi\rangle \\ &= \langle\varphi|\,\hat{O}|\hat{I}\,|\varphi\rangle \\ &= \langle\varphi|\,\hat{O}\,|\varphi\rangle \end{aligned} \tag{2.91}$$

Note that this treatment of expectation value depends on the Born rule. Thus, we cannot consider $\langle\varphi|\hat{O}|\varphi\rangle$ as an expectation value (or average result) if we are attempting to derive the Born rule, as that would be circular reasoning.

Recall that a projector can be interpreted as asking a "yes-no" or "1-0" question. Take, for example, a binary basis of $\{|0\rangle, |1\rangle\}$. Let's form a ket from this that has components in both directions of the 2-D basis. The lengths of the components will have to be less than unity, since our final ket must be of unit length (since our kets are always normalized by convention). Assume the length of the $|0\rangle$ component is 1/2. The length, then, of the other component must be determined by the Pythagorean theorem [78, Bk.1, Pr.47], since the basis vectors are orthogonal:

$$\begin{aligned} 1^2 &= ||0\rangle|^2 + ||1\rangle|^2 \\ ||1\rangle|^2 &= 1^2 - \left(\frac{1}{2}\right)^2 \\ ||1\rangle| &= \frac{\sqrt{3}}{2} \end{aligned} \tag{2.92}$$

and our resultant ket must be (to maintain normalization):

$$|\psi\rangle = \frac{1}{2}|0\rangle + \frac{\sqrt{3}}{2}|1\rangle \tag{2.93}$$

Now apply the projector $[0] = |0\rangle\langle 0|$, which asks "Do you contain a 0?" The question can be posed to each component distributively, and the answer will clearly be "yes" for one and "no" for the other:

$$\begin{aligned} [0]\,|\psi\rangle &= \frac{1}{2}|0\rangle\langle 0|0\rangle + \frac{\sqrt{3}}{2}|0\rangle\langle 0|1\rangle \\ &= \frac{1}{2}|0\rangle + \frac{\sqrt{3}}{2}0 \\ &= \frac{1}{2}|0\rangle + 0|1\rangle \end{aligned} \tag{2.94}$$

where the $|0\rangle$ is the basis state indexed by 0, and the 0 itself is the null vector[11]. Taken as a result on its own, this ket can just be normalized to $|0\rangle$, which corresponds to an answer of 1 or "yes" to the [0] question $(1|0\rangle + 0|1\rangle)$, or 0 or "no" for the [1] question $(0|0\rangle + 1|1\rangle)$. In quantum mechanics, this normalized projection might correspond to the collapse of the wavefunction after a measurement. But as an *intermediate* value in an expectation value calculation we cannot, of course, normalize

[11] One can see here why it can be confusing to use $|0\rangle$ for the null vector.

yet:

$$\begin{aligned}\langle\psi|0\rangle\langle 0|\psi\rangle &= \langle\psi|\tfrac{1}{2}|0\rangle = \frac{1}{2}\langle\psi|0\rangle \\ &= \frac{1}{2}\left(\frac{1}{2}\langle 0| + \frac{\sqrt{3}}{2}\langle 1|\right)|0\rangle \\ &= \frac{1}{2}\left(\frac{1}{2}\langle 0|0\rangle + \frac{\sqrt{3}}{2}\langle 1|0\rangle\right) \\ &= \frac{1}{2}\left(\frac{1}{2}+0\right) \\ &= \frac{1}{4}\end{aligned} \tag{2.95}$$

which is the same result as simply multiplying the two inner products:

$$\langle\psi|0\rangle\langle 0|\psi\rangle = \frac{1}{2}^{*}\frac{1}{2} = \frac{1}{4} \tag{2.96}$$

and we see that, as expected, the expectation value associated with the projector is the probability of that projector's state, which is the square of the amplitude for that state:

$$\langle\psi\,[k]\,\psi\rangle = \langle\psi|k\rangle\langle k|\psi\rangle = |\langle k|\psi\rangle|^2 = |c_k|^2 \tag{2.97}$$

2.3 The Postulates of Quantum Mechanics

2.3.1 Introduction

Quantum mechanics is derived from five fundamental postulates. These can be stated in a fairly general way that is not tied to the particularities of any actual physical measurement situation. In this sense, "quantum mechanics" is a general calculus that can take the form of a number of particular physical theories involving particular types of observables (such as electrons or photons and so on). Quantum electrodynamics is thus a particular quantum mechanical theory involving electrons and photons. But quantum mechanics itself—as defined by the five postulates—does not yield any particular particle types, nor even state that the observables are (or even can be) "particles".

The five postulates are therefore expressed sometimes in more general, and sometimes more particular forms, depending on ones's goals (and sometimes there may be somewhat fewer or more than five). Given my interpretational and foundational goals, my version of the postulates will be quite general. And it is this formulation of the postulates that I will try to show later (in Ch. 8) to be derivable, at least in outline, from more basic—and non-empirical—philosophical principles.

I will divide the postulates into two basic types: analytic and synthetic. The analytic postulates define (precisely and analytically) what a quantum state (or "wavefunction") is, while the synthetic postulates define (imprecisely and synthetically) what a "measurement" or "observation" is.

2.3.2 The Analytic Postulates

The analytic postulates (#1-2) define quantum states (#1) and their evolution (#2) in purely analytic terms.

Postulate. (1) The Superposition Principle: *The "state" of a (physical) system or subsystem Ψ is represented by a unit-norm vector $|\psi\rangle$ (the "wavefunction") in Hilbert space $\mathcal{H}$ (where any non-zero vector can simply be scaled to unit norm to obtain the "state" that it represents).*

We call this state a "superposition" of states $\{|\psi_k\rangle\}$ when the $|\psi_k\rangle$ are mutually orthogonal and $|\psi\rangle$ can be expressed as a linear combination of them:

$$|\psi\rangle = \sum_k a_k |\psi_k\rangle = \sum_k |\psi_k\rangle \langle\psi_k|\psi\rangle \tag{2.98}$$

where $a_k = \langle\psi_k|\psi\rangle$ is a complex scalar, called the "amplitude" for state $|\psi_k\rangle$ given state $|\psi\rangle$. When $\{|\psi_k\rangle\}$ spans Hilbert space $\mathcal{H}$, it is a "basis" for that space.

Postulate. (2) Unitary Evolution: *Any evolution or transformation of a (physical) state $|\psi\rangle$ must be describable with a unitary operator, $\hat{U}$.*

This implies that amplitudes are preserved:

$$\left\langle \hat{U}\varphi \middle| \hat{U}\psi \right\rangle = \langle\varphi|\psi\rangle \tag{2.99}$$

2.3.3 The Synthetic Postulates

The synthetic postulates (#3-5) define (loosely) what a "measurement" or "observation" is, by relating our synthetic knowledge about ordinary observations to the analytic model provided by the first two postulates. Their role is not to describe the dynamics of the system, or, for that matter, to provide any kind of analysis at all of what a quantum system *is*. Their role, rather, is to interface the analytic description already provided to *empirical* phenomena.[12]

[12] As alluded to in the previous section, this is not some weird dodge; it is how science works... all so-called analytic or mathematical models of nature that are employed in science still need to be interfaced to empirical phenomena, and this interfacing will always involve some kind of synthetic hand-waving at some point. Thus, even though science seeks, so far as possible, to provide analytic models of nature, no scientific theory is "purely" analytic. This is why, although I do not believe in dismissing the Born rule objection entirely out of hand with the double-standard or bemusement responses, I do believe there is something to these responses. The Born rule objectors *do* seem to be requiring, at times, an unusual degree of *analytic* justification for the "interface" between theory and phenomena, given that scientific models cannot in general be expected to provide this kind of total "self-interfacing" to the empirical realm (but, then, oddly at other times, they seem to want to hold everything to the purely synthetic standard of branch-counting, for which they see no need to present any justification at all). The reason I do not believe in dismissing the Born rule objection out of hand, however, is that, while it may be true that scientific models do not "self-interface", the counter-intuitive context of quantum mechanics means that it is not as intuitively clear how to perform this interfacing, compared to classical physics, and much scientific progress can be made by the attempt to

Postulate. (3) Observability Postulate: *For each measurable property, or observable $\mathcal{O}$ of a system or subsystem $|\psi\rangle$, there is identified a corresponding linear Hermitian operator $\hat{O}$, for which the possible measurement outcomes are the (real) eigenvalues $\{o_k\}$ of members of an orthonormal "preferred" basis of "observable eigenstates" of $\hat{O}$, $\{|k\rangle\}$.*

Postulate. (4) Collapse Postulate: *The result, for a particular observer, of any act of observation will be one (and only one) of the eigenvalues for the preferred basis, and after the observation, the observed system or subsystem will be left, for that observer, in the eigenstate associated with the observed eigenvalue (or, if the preferred basis is degenerate, the system will be left in the projection of $|\psi\rangle$ onto the subspace defined by the eigenstates with that eigenvalue).*

Postulate. (5) Probability Postulate *(Born rule)*: *The probability of observing the eigenvalue that leaves the system in state $|\psi'\rangle$ is proportional to the norm-squared of the amplitude for $|\psi'\rangle$ in $|\psi\rangle$:*

$$p(|\psi'\rangle) \propto |\langle\psi'|\psi\rangle|^2 \tag{2.100}$$

2.3.4 Discussion

Note that by calling postulates #1 and #2 "analytic", we are not really claiming that they are purely and absolutely analytic statements, in the strict philosophical sense, since the analytic expressions they provide *are*, after all, meant to model the external physical world, and so are still empirical in that sense (and all postulates or axioms for any theory or system are synthetic to the extent that they are presumed without logical justification). However, the first two postulates are claiming that physical systems can be modelled by certain purely analytic expressions, which *can* be precisely described without reference to anything external or synthetic. In that sense, their synthetic content is very minimal, consisting (perhaps) only in the parenthetical inclusion of the word "physical" in each postulate. The synthetic postulates, on the other hand, barely give any analytic expression at all to what they are modelling—since they are modelling *observation*, and no analytic model of observation is even attempted here. The synthetic postulates thus are content to simply *interface* an ill-defined synthetic notion of "observation" with the more precise analytic model already provided in the first two postulates.

---

further analyze the analytic-synthetic boundary line, whether we expect to completely erase it or not. And even if we decide that we can never entirely erase this line, this is no justification for assuming that we cannot continue to push the line further and further into the analytic and away from the synthetic. But this still does not justify, of course, the view that the MWI *must* provide a 100% analytic accounting of probability or be dead in the water. Not all Born rule objectors are necessarily guilty of this.

The fact that these last three postulates are inherently synthetic—and that they are so precisely *because* they are about observation—should give us an early warning that we should not expect any of these synthetic postulates to be formally derivable from the first two postulates. It should also hint, however, that we might perhaps find deeper insight into the synthetic postulates by finding a more analytic model of the presently ill-defined process called observation or measurement.

If we were to ignore the synthetic nature of the last three postulates, and attempt to interpret them as analytic models, by making them descriptive of quantum systems themselves, this would put us in direct contradiction with the analytic postulates.[13] This is, at least, clearly so with the collapse (#4) and probability (#5) postulates, which would directly violate the linearity of the analytic postulates (*if* taken as analytic themselves). Observability (postulate #3) would *not* directly create such a violation on its own, since it simply states what it means for a property to be "observable", without making any claims about what actually happens when an observation is made. On the other hand, it is not clear how to make sense of such a notion of observability without introducing something like the nonlinearity inherent in the other two synthetic postulates. It is this inherent tension between the synthetic and analytic postulates that generates what is known as the "measurement problem".

2.4 The Wavefunction

2.4.1 The wave equation

2.4.1.1 The basic (spatial) wave equation

We can write the wave equation for a simple one-dimensional spatial wave, ψ, as an amplitude a times a phase factor:[14]

$$\begin{aligned} \psi(x) &= a e^{i\dot{k}_x x} \\ &= a e^{i\dot{k}x} = a e^{ikx\tau} \end{aligned} \tag{2.101}$$

[13] I do not mean to read the synthetic postulates as ruling out the possibility that they can be given an analytic interpretation, but taking such a stance would clearly be a particular interpretation of the postulates, and hence something that would go beyond the postulates themselves (and it would certainly not be standard).

[14] I commit a (fairly common) abuse of notation here, and elsewhere in this dissertation. Strictly speaking, $\psi(x)$ and $\psi(k)$ should be written as distinct functions, possibly as $\psi_x(x)$ and $\psi_k(k)$. They are clearly different functions, even though they represent the same wavefunction ψ. However, leaving out the subscripts simplifies things without ambiguity, so long as it is clear from the argument whether it represents a momentum (spatial frequency) or a position. In Dirac notation, the same abuse occurs when we write $\langle x|\psi\rangle$ instead of writing, for instance, something like $\langle x_i|\psi\rangle$. The reason for the abuse is that we usually want to use x (not i) as the index for the position basis, which leads to the pedantic $\langle x_x|\psi\rangle$, with projectors $|x_x\rangle\langle x_x|$. We need to drop the abuse, however, when specifying an actual index, unless the intended basis is clear from context. Thus, we write something like $\langle x_1|\psi\rangle$ or $\langle b_1^x|\psi\rangle$ or even $\langle x=1|\psi\rangle$, instead of $\langle 1|\psi\rangle$, unless it is clear from the context that we are working in position-space.

in terms of spatial frequency

$$\begin{aligned} \dot{k} &= k\tau \\ k &= \frac{1}{\lambda} \end{aligned} \tag{2.102}$$

where

k_x is spatial frequency (wavenumber, momentum) in the x direction,
$\dot{k}_x$ is angular spatial frequency (angular wavenumber) in the x direction,
λ_x is wavelength (spatial period) in the x-direction, and
τ is angular unity (the circumference of the unit circle).

We will use x, y, and z subscripts to indicate spatial (and t for temporal) direction, but freely omit such subscripts when direction is understood.

The wave equation, as written above, represents a two-dimensional structure, defining a simple one-dimensional wave displaying a very basic periodic structure (there is no "dynamics" here, since there is no time). We will call this the "basic" (spatial) wavefunction, as it represents the simplest, prototypical case (it is so simple, we have not yet even added a temporal dimension).

2.4.1.2 The basic spatiotemporal wave equation

We can modify the basic (spatial) wavefunction to yield a slightly more complex "basic spatiotemporal wavefunction":

$$\psi(x,t) = ae^{i(\dot{k}x - \dot{f}t)} \tag{2.103}$$

$$= ae^{i(kx - ft)\tau} \tag{2.104}$$

in terms of temporal frequency

$$\dot{f} = f\tau \tag{2.105}$$

$$f = \frac{1}{T} \tag{2.106}$$

where

$f = k_t$ is ordinary (temporal) frequency,
$\dot{f}$ is angular frequency, and
$T = \lambda_t$ is period (or temporal wavelength).

Note that when adding a time dimension to the wave equation, we *subtract* the temporal frequency (whereas we *add* spatial frequencies).

2.4.1.3 The basic higher-dimensional wave equation

More complex periodic structures can also be built with the same basic wave equation. Adding just another spatial dimension, for instance, gives us

$$\psi(x,y) = ae^{i\dot{k}_x x}e^{i\dot{k}_y y} \quad = ae^{i(\dot{k}_x x+\dot{k}_y y)} \tag{2.107}$$

$$= ae^{i(k_x x+k_y y)\tau} \tag{2.108}$$

We can generalize to arbitrary dimensions, and just replace k, $\dot{k}$ and x with vectors $\mathbf{k}$, $\dot{\mathbf{k}}$ and $\mathbf{r}$:

$$\begin{aligned} \psi(\mathbf{r}) &= ae^{i\dot{\mathbf{k}}\mathbf{r}} \\ \psi(\mathbf{r},t) &= ae^{i(\dot{\mathbf{k}}\mathbf{r}-\dot{f}t)} \end{aligned} \tag{2.109}$$

With a three-dimensional $\mathbf{r} = (x,y,z)$, and one dimension of time t, we get the form that is most applicable to the real world:

$$\psi(\mathbf{r},t) = \psi(x,y,z,t) \tag{2.110}$$

$$= ae^{i(\dot{\mathbf{k}}\mathbf{r}-\dot{f}t)} = ae^{i(\dot{k}_x x+\dot{k}_y y+\dot{k}_z z-\dot{f}t)} \tag{2.111}$$

$$= ae^{i(\mathbf{k}\mathbf{r}-ft)\tau} \tag{2.112}$$

For simplicity, I will generally stick with two dimensions, one spatial (x) and one temporal (t). Note that while the wavenumbers k sum across all spatial dimensions, the analog for time, f, is *subtracted* from this total.

2.4.2 General (solved) Schrödinger's equation

By the superposition principle, we can take our wavefunction ψ to be a superposition of waves. The most general approach is to consider it a superposition of basic plane waves, as described by the wave equation. This superposition can be expressed in various "bases", or "domains". For instance, we could take it to be a superposition across all possible spatial positions, x:

$$\psi(k) = \int_{-\infty}^{\infty} \psi_x(k)dx \tag{2.113}$$

or as a superposition across the spectrum of spatial frequencies, k:

$$\psi(x) = \int_{-\infty}^{\infty} \psi_k(x)dk \tag{2.114}$$

where

$$\psi_k(x) = \psi(x) \dashrightarrow \psi(k) = \psi(k)\, e^{ikx\tau}$$
$$\psi_x(k) = \psi(k) \dashrightarrow \psi(x) = \psi(x)\, e^{-ikx\tau}$$

(We are starting here with just one dimension of space; we will add time shortly.)

These are the unitary "Fourier transforms". The former is called the "Fourier analysis", and the latter the "Fourier synthesis," of wavefunction ψ.[15] If the spectrum of positions is finite and discrete at resolution N, then $\{|x\rangle : x = 1 \cdots N\}$ is an orthonormal set of basis vectors (defining the "spatial domain") such that

$$|x\rangle \quad = \quad \begin{bmatrix} 1 \\ \vdots \\ e^{i\frac{k}{R}x\tau} \\ \vdots \\ e^{i\frac{R-1}{R}x\tau} \end{bmatrix} \tag{2.115}$$

with x values normalized over N. Likewise, the inverse transform uses the "frequency domain", defined by basis $\{|k\rangle : k = 1 \cdots R\}$:

$$|k\rangle = \begin{bmatrix} 1 \\ \vdots \\ e^{-ik\frac{x}{N}\tau} \\ \vdots \\ e^{-ik\frac{N-1}{N}\tau} \end{bmatrix} \tag{2.116}$$

with k values normalized over R, and where each basis set is an orthogonal basis spanning the set of N and R-dimensional complex vectors, respectively. Discrete, finite basis sets are assumed above, but the bases can be expressed in continuous terms, as well. Expressing the forward and inverse transforms in discrete and continuous functional notation, and in Dirac notation, we have:

$$\langle k \mid \psi \rangle = \sum_x \langle k \mid x \rangle \langle x \mid \psi \rangle \qquad \psi(k) = \frac{1}{R} \sum_{x=0}^{N-1} \psi(x) e^{-ik\frac{x}{N}\tau} \qquad \psi(k) = \int_{-\infty}^{\infty} \psi(x) e^{-ikx\tau} dx$$
$$\langle x \mid \psi \rangle = \sum_k \langle x \mid k \rangle \langle k \mid \psi \rangle \qquad \psi(x) = \frac{1}{N} \sum_{k=0}^{R-1} \psi(k) e^{i\frac{k}{R}x\tau} \qquad \psi(x) = \int_{-\infty}^{\infty} \psi(k) e^{ikx\tau} dk \tag{2.117}$$

[15] Mathematically, both operations really just transform the data from a representation in one basis to a representation in another, integrating in the same basic way, so it thus is somewhat arbitrary which we consider to be "analytic" and which "synthetic", and both operations are sometimes simply lumped together as "Fourier analysis". However, because it is more typical for us to think of spatial information as "given" or phenomenal, and frequency-based information as something inferred mathematically *from* this data, the spatial-to-frequency direction is most often chosen as the "synthetic" direction, and we then talk about re-constructing or synthesizing the original spatial data out of this frequency-based analysis, or even purely artificial frequency values. (This is why a music "synthesizer" is never called a music "analyzer", and a Fourier "analysis" of recorded audio is never called "synthesis".)

When we add a temporal dimension t, these become

$$\langle k \mid \psi \rangle = \sum_x \langle k \mid x, t_i \rangle \langle x, t_i \mid \psi \rangle \quad \psi(k) = \frac{1}{R} \sum_{x=0}^{N-1} \psi(x, t_i) e^{-ik\frac{x}{N}\tau} \qquad \psi(k) = \int_{-\infty}^{\infty} \psi(x, t_i) e^{-ikx\tau} dx$$

$$\langle x, t \mid \psi \rangle = \sum_k \langle x, t \mid k \rangle \langle k \mid \psi \rangle \qquad \psi(x, t) = \frac{1}{N} \sum_{k=0}^{R-1} \psi(k) e^{i\left(\frac{k}{R}x - ft\right)\tau} \quad \psi(x, t) = \int_{-\infty}^{\infty} \psi(k) e^{i(kx - ft)\tau} dk \tag{2.118}$$

where t_i is any arbitrary value for time t, but we will informally call it "initial time", with a typical choice being $t_i = 0$.

This is the general solved version of Schrödinger's equation (generalizable to three spatial dimensions by replacing x and k with $\mathbf{r}$ and $\mathbf{k}$). The discrete transform can be used to actually calculate the wavefunction for some final point in time t_f, given the initial conditions at time t_i, without the need to step-wise calculate the dynamics for the intermediate times.[16] Just follow this procedure (for speed, you might want to use one of the highly efficient "fast Fourier transform" (FFT) algorithms that are available for computing DFTs):

1. Use the forward discrete Fourier transform (DFT) on the initial conditions ($t = t_i$), to get a frequency-based representation of the initial conditions:

$$\psi(k) = \frac{1}{R} \sum_{x=0}^{N-1} \psi(x, t_i) e^{-ik\frac{x}{N}\tau} \tag{2.119}$$

   for $k = 0 \cdots (R-1)$.

2. Transform back (after an optional phase shift) to the spatiotemporal domain with the inverse DFT for the desired final time ($t = t_f$):

$$\psi(x, t_f) = \frac{1}{N} \sum_{k=0}^{R-1} \psi(k) e^{i\left(\frac{k}{R}x - ft_f\right)\tau} \tag{2.120}$$

I use the discrete transform to illustrate this, since I am emphasizing our ability to use the Fourier transform to actually calculate—the continuous version is generally uncomputable, since continuous variables contain an infinite amount of information. So even putting aside the fact that there is absolutely no evidence for the existence of continuities in nature, and assuming we are actually measuring a real, physically continuous system, one can still do so only to some coarseness of scale, and we will thus still be calculating a *discrete* Fourier transform, in practice (which will be an approximation, even if a very accurate and detailed one).

[16] Initial time t_i does not have to be 0, and final time t_f does not have to be in the future, since all the information contained in the wavefunction is contained in any single time-slice of it. While it may be most typical to want to compute the state of the wavefunction at a future time given some initial state, the transform can equally be used to calculate the initial conditions, given the state of the system right now. In general, we can calculate the state at any arbitrary point, given the state at any other arbitrary point.

Note that the inverse transform still sums or integrates over *spatial* frequencies only, even though we have now added in a time dimension. So even though time makes its appearance here, these are still technically *spatial* forward and inverse transforms even though performed on a spatio-*temporal* wavefunction—space and time are not being treated equally here.[17]

Finally, it is worth noting that, mathematically, there are many more possible bases, and we could have chosen a completely different starting point than standard plane waves. We could have chosen any number of more complicated waveforms as a basis. In fact, we can still do so, and we can create transforms that take us from momentum or position bases to these other bases. Do you like piano music? Use the piano waveform as your basis, and you will have a system of transforms analogous to that above. However, this would add unnecessary complexity to the system unless we were working in a domain where piano music was justifiably fundamental (which is, of course, possible). However, given the simplicity and elegance of the basic wave equation, the momentum (k) and position (x) bases are very simple and natural bases to work with, and have special importance in the practice of quantum mechanics.

2.4.3 Dynamical Schrödinger's equations

We can differentiate the basic wave equation, either with respect to space (x) or time (t), to yield the basic dynamical (unsolved) versions of Schrödinger's equation.

2.4.3.1 The basic time-independent Schrödinger's equation

Recalling either the spatial wave equation,

$$|\psi\rangle = ae^{i\dot{k}x} \tag{2.121}$$

or the spatiotemporal version,

$$|\psi\rangle = ae^{i(\dot{k}x-\dot{f}t)} \tag{2.122}$$

[17] One could, of course, also construct a more general, truly *spatiotemporal* transform, by simply treating the time dimension as a spatial dimension with a negative wavenumber, so $k_t = -f$, and using the purely spatial version of the transforms. One reason for *not* doing this, other than the mere convenience of making time distinct, is that we do not actually add any information to our representation of the wavefunction at time t_i by then integrating over all time as well. This would be adding an enormous amount of computation that would be completely unnecessary. The transform between the spatial and frequency domains is unitary, and therefore information-preserving. The transform from time t_i to t_f is like-wise unitary and information-preserving. However, there is no symmetry transform from spatial dimension x to y, nor from spatial position x_i to x_f. These would be projections, not symmetry transformations, and information would be lost. Thus, we *must* in general integrate over all the dimensions of space, but *not* necessarily of time, in order to have a proper unitary Fourier transform. Hence, generalizing to a fully "spatiotemporal" transform would be redundant, as would the creation of a purely "temporal" version by summing over all times but *not* spatial positions.

in either case, if we differentiate with respect to x (independent of t, if it exists) we get

$$\frac{\partial}{\partial x}|\psi\rangle = i\dot{k}|\psi\rangle \tag{2.123}$$

and we have the basic time-independent form of Schrödinger's equation.

2.4.3.2 The general time-independent Schrödinger's equation

To allow applications to more specific situations, with more complex Hamiltonians, we can simply replace $\dot{k}$ in (2.123) with a general Hamiltonian operator $\hat{H}$ for total energy (in whatever form we choose to express it):

$$\frac{\partial}{\partial x}|\psi\rangle = i\hat{H}|\psi\rangle \tag{2.124}$$

This is the most general form of the time-independent Schrödinger's equation, and can be applied to all kinds of specific situations, to yield more specific forms of Schrödinger's equation. There will not necessarily be a precise frequency value here, like k in the basic equation, unless the system is an eigenket of the Hamiltonian, as is the case for the standing waves we see in atomic orbitals, where we then get the following (specific) solved version of Schrödinger's equation

$$\hat{H}|\psi\rangle = E_k|\psi\rangle \tag{2.125}$$

where k now indexes the energy level of the orbital.

Another common application is to use the standard classical equations for mechanical (kinetic and potential) energy to create the Hamiltonian (see Appendix A).

2.4.3.3 The basic time-dependent Schrödinger's equation

To calculate a time-dependent version of Schrödinger's equation, we differentiate the wave equation

$$|\psi\rangle = ae^{i(\dot{k}x - \dot{f}t)} \tag{2.126}$$

but, this time, with respect to time t:

$$\frac{\partial}{\partial t}|\psi\rangle = -i\dot{f}|\psi\rangle \tag{2.127}$$

2.4.3.4 The general time-dependent Schrödinger's equation

As before, we find the general case by replacing the basic energy quantity $\dot{f}$ with a generalized Hamiltonian operator, $\hat{H}$:

$$\begin{aligned} \frac{\partial}{\partial t}|\psi\rangle &= -i\hat{H}|\psi\rangle \\ i\frac{\partial}{\partial t}|\psi\rangle &= \hat{H}|\psi\rangle \end{aligned} \tag{2.128}$$

This is the most general form of the time-dependent Schrödinger's equation. Like the time-independent equation, it too can be applied to all kinds of specific situations, such as classical mechanical energy (see Appendix A).

2.5 The Measurement Problem

2.5.1 Measurement

We can view a "measurement" as an interaction between a system S and an apparatus A. The subsystem S is represented by a vector $|\sigma\rangle$ in Hilbert space H_S that can be decomposed according to a set of basis vectors $\{|b_n\rangle\}$, as a linear combination of the members of the basis set

$$|\sigma\rangle = \sum_k s_k |\sigma_k\rangle \tag{2.129}$$

Vector $|\sigma\rangle$ can be described in terms of any number of such basis sets. One such basis set will be the set that consists solely of $|\sigma\rangle$ itself: $\{|\sigma\rangle\}$. In this case, the weighted sum is trivial

$$|\sigma\rangle = s |\sigma\rangle = |\sigma\rangle \tag{2.130}$$

The constant s could be any complex number, since a single vector multiplied by any scalar represents the same quantum state (so we might as well consider that $s = 1$).

The measuring apparatus A is really just another quantum system; it is likewise represented by a vector $|\alpha\rangle$ in Hilbert space H_A, described by basis vectors $\{|\alpha_k\rangle\}$

$$|\alpha\rangle = \sum_k a_k |\alpha_n\rangle \tag{2.131}$$

Since A is just another quantum system, it can *analytically* be described, like S, by any number of basis sets, of which one is the set $\{|\alpha_k\rangle\}$ containing only the vector $|\alpha\rangle$ itself. However, in practice, when A really does represent an actual measuring apparatus—such as a gauge with a pointer readout—there will be a synthetically "preferred" basis set whose members correspond to the actually discernible pointer states of the device. Assume that $|\alpha_k\rangle$ corresponds to a macroscopically distinguishable pointer position that corresponds to the k-th eigenvalue of the operator, which measures S as being in state $|s_n\rangle$.

Assume S is in state $|\psi\rangle = \sum_n c_n |s_n\rangle$, which is a superposition in the $\{|s_k\rangle\}$ basis. Assume that apparatus A is in an initial "ready" state $|a_{ready}\rangle$. Call the total system SA (system S and apparatus A). If S is represented in Hilbert space $\mathcal{H}_\mathcal{S}$ and A in Hilbert space $\mathcal{H}_\mathcal{A}$, then SA can be represented in the Hilbert tensor product space $H_S \otimes H_A$. This is the synthetically "preferred factoring" of the Hilbert space.

Thus the total system state of S being in state $\sum_n c_n |s_n\rangle$, and A being in state $|a_{ready}\rangle$, is represented by the "tensor product"

$$\left(\sum_n c_n |s_n\rangle\right) \otimes |a_{ready}\rangle \tag{2.132}$$

Definition 2.17. The *unitary evolution* arrow $\Longrightarrow$ (read "evolves to") will represent the unitary evolution of the wavefunction. S

$$|\psi_t\rangle \Longrightarrow |\psi_{t+1}\rangle \tag{2.133}$$

means that state $|\psi_t\rangle$ evolves unitarily to state $|\psi_{t+1}\rangle = \hat{U} |\psi\rangle$, where $\hat{U}$ is a unitary operator.

We now use the unitary arrow to represent the unitary evolution from the above state of SA (in which we assume S and A are uncorrelated, with zero mutual information) to a state wherein S and A are correlated, A having obtained the relevant information about S to be said to have measured it

$$\left(\sum_n c_n |s_n\rangle\right) \otimes |a_{ready}\rangle \Longrightarrow \sum_n c_n \otimes |s_n\rangle \otimes |a_n\rangle \tag{2.134}$$

To keep things compact, the tensor product symbol can be omitted

$$\left(\sum_n c_n |s_n\rangle\right) |a_{ready}\rangle \Longrightarrow \sum_n c_n |s_n\rangle |a_n\rangle \tag{2.135}$$

To represent the collapse necessary for an actual measurement to have occurred (ignoring here whether this is a synthetic matter of perspective, or a real mechanical change), we use a different arrow.

The *collapse* arrow $\rightarrowtail$ (read "collapses to") will represent the wavefunction collapse (regardless of whether one considers it to be a real physical mechanism, or merely a matter of taking a certain perspective). So

$$|\psi\rangle \rightarrowtail |s_k\rangle$$

means that state $|\psi\rangle$ collapses to one of n incompatible components, $|s_1\rangle, \ldots, |s_n\rangle$, each of which represents a different possible measurement/observation result, where $|\psi\rangle = \sum_i c_i |s_i\rangle$ is a superposition of all these results. This corresponds to a projection onto the subspace defined by the possible outcome states, the collection of which is the preferred basis of the measurement. When we say that the states to the left of the arrow, in superposition, are "incompatible" with each other, we mean that they are states that are never observed as true of the world at the same time.

So the collapse of SA can be represented a

$$\sum_n c_n |s_n\rangle |a_n\rangle \rightarrowtail |s_k\rangle |a_k\rangle \tag{2.136}$$

where k is some arbitrary (random) number from 1 to n.

The entire measurement process can be summarized a

$$\left(\sum_n c_n \left|s_n\right\rangle\right)\left|a_{ready}\right\rangle \Longrightarrow \sum_n c_n \left|s_n\right\rangle\left|a_n\right\rangle \rightarrowtail \left|s_k\right\rangle\left|a_k\right\rangle \tag{2.137}$$

To indicate a collapse from $|\psi\rangle$ specifically to the k-th basis state, we can annotate the arrow like so

$$|\psi\rangle \overset{k}{\rightarrowtail} |s_k\rangle \tag{2.138}$$

To indicate that k picks out the k-th component in a particular basis, we can indicate below the arrow the basis from which we are choosing our selected state

$$|\psi\rangle \underset{\{|s_i\rangle\}}{\overset{k}{\rightarrowtail}} |s_k\rangle \tag{2.139}$$

More than one index may be specified above the arrow, if the selected state is a combination of more than one basis state. The annotation need not actually index a basis, however; it can specify or summarize the selected branch in any way we wish, including the use of descriptive English phrases, so long as the annotation above the arrow (if there is one) describes which state is selected, while the one below the arrow (if there is one) describes the preferred basis for the measurement. (When the annotations are less than precise, the intended meaning should be clear from the context.)

The annotation "*eg*" above the arrow will mean that we are being shown a state, to the right of the arrow, that is merely an example of one possible state the system might collapse to (so there is nothing special about selecting that particular state)

$$|\psi\rangle \overset{eg}{\rightarrowtail} |s_k\rangle \tag{2.140}$$

Definition 2.18. The *uncollapse* arrow $\longrightarrow\!\!\!\!\prec$ (read "uncollapses (or merges) into") is the time-reversal of the collapse process, where n incompatible states merge into a superposition of states

$$\left|s_1\right\rangle, \ldots, \left|s_n\right\rangle \longrightarrow\!\!\!\!\prec |\psi\rangle \tag{2.141}$$

where $|\psi\rangle = \sum_{i=1}^{n} c_i \left|s_i\right\rangle$, so that $|\psi\rangle$ is a superposition of all states listed to the left of the arrow.

When we say the states to the left of the arrow are "incompatible" with each other, we mean that they are states that are never observed as true of the world at the same time. Remember that the states listed to the left of the arrow are *not* in superposition, they are incompatible states of the system merging into a superposition to the right of the arrow. From the MWI point of view, this would mean that we had multiple worlds merging into one world (the time reverse of one world bifurcating into multiple worlds, which is the usual MWI order of things). From the collapse-as-real-mechanism point of view, this would mean something like a superposition of incompatible results appearing, discontinuously, where there was only a single definite result before (clearly, the whole idea makes less sense for the real-collapse interpretations).

As with collapse, we can indicate which states to merge above the arrow

$$|s_i\rangle\,, |s_j\rangle \xrightarrow{i,j} \!\!\!\!\!\!\succ\; |\psi\rangle \tag{2.142}$$

If there is one state we are considering the pre-existing state, we can show it alone to the left of the arrow, and show the other state or states that it is to merge with above the arrow

$$|s_i\rangle \xrightarrow{j} \!\!\!\!\!\!\succ\; |\psi\rangle \tag{2.143}$$

We can also indicate a basis below the arrow, as with the collapse annotation. However, keep in mind that, prior to a collapse, there is no *a priori* reason to prefer one basis over another. We generally speak of a "problem of the preferred basis" because it is more or less mysterious why, when collapse occurs, a certain basis is always preferred. With uncollapse, however, we *start* with incompatible states, already independent of one another, that then merge them into superposition, so we already have an *a priori* preference for a basis (or at least certain bases) *before* the uncollapse occurs. Once they have merged, however, there is no longer any *a priori* reason to prefer the old basis.

I should explain why I introduce this third arrow whose inclusion here might seem puzzling and perhaps pointless. Indeed, since we do not observe time-reversed collapses of the wavefunction in the real world (meaning not in the forward direction of time) this third arrow is really only useful for analyzing thought experiments, and other unrealistic situations, in which the time-reverse of collapse is being taken seriously, for whatever reason. This is rarely necessary for interpretations where the collapse is considered a real physical process. However, for "no-collapse" interpretations like the MWI, where the collapse is viewed as merely perspectival—and the unitary evolution of the wavefunction is the *only* true mechanism involved—one must at least admit the theoretical possibility of the reverse of collapse taking place, since the wavefunction mechanics, on their own, are time-symmetric.[18] This does not make it any more strange that the arrow only goes one way in practice than it is strange that there is an arrow of time in Newtonian mechanics, which is also time-symmetric.

To clarify this a bit, consider a situation in which "uncollapse" might take place. We have a superposition of spin-up $|+\rangle$ and spin-down $|-\rangle$ states for a spin-$1/2$ system

$$|\psi\rangle = \frac{1}{\sqrt{2}}\,|+\rangle + \frac{1}{\sqrt{2}}\,|-\rangle \tag{2.144}$$

It is possible to talk about very simple systems "collapsing", since we can "take the perspective of" a single particle or coherent system of particles (usually microscopic, but definitely not a real, conscious

[18] The time-reversal of the unitary mechanics of Schrödinger's equation is actually anti-unitary, rather than unitary, which might make it seem that time-reversal is not symmetric. However, anti-unitary operators are also (like unitary operators) symmetry operators. So time-reversing a unitary transformation may not itself be unitary, but it is still symmetrical and there is hence no loss of information or structure.

observer). For instance, we could choose to talk about a particle's colliding with another particle and "collapsing" its spin superposition from $\frac{1}{\sqrt{2}}|+\rangle + \frac{1}{\sqrt{2}}|-\rangle$ into the specific state $|+\rangle$. However, such "collapse" of microscopic systems is not observed. Collapse is only ever observed (and even then, not directly) at the scale of macroscopic observers (which is why it is unclear whether the phenomenon is a real mechanism or a matter of perspective). Nonetheless, it can still be convenient, especially in thought experiments, to sometimes "take the perspective of" a single particle or microscopic system (as if it were an observer), and use the language of collapse. This is permitted, so long as it is clear from the context the sense in which "collapse" is intended to be taken. Usually, however, we will use the collapse symbol to represent actual observation or measurement.

The unitary evolution (before a collapse is considered), we will call a "pre-measurement", while the result after the collapse, we will call the "post-measurement". However, it should be emphasized, that this language is a bit misleading, and is not meant to commit us to the idea that the post-measurement actually occurs *after* the pre-measurement in time. After all, if the collapse is just a matter of perspective, then the given collapse transformation ($\rightarrowtail$) is not an evolution in time at all, nor any other kind of mechanism. But even if one takes the perspective that the collapse *is* a real mechanism, the use of "pre" and "post" prefixes are still not meant to indicate that one necessarily occurs before the other. There could, after all, be a real collapse mechanism that takes place over the course of the otherwise-unitary evolution, and not strictly after it has already completed. Nonetheless, our notation will make use of the linguistic convenience of talking as if the unitary evolution takes place, and then an instantaneous collapse occurs, in order to emphasize that the process described by Eq. 2.1 does not suffice to directly conclude that a measurement has actually been completed (just keep in mind that this is a mere convenience, not a claim about what really happens).

Note, also, that the reverse collapse notation represents something (re-merging or recoherence) which is often said to be impossible. In fact, however, there is no reason, from the formalism of the quantum wavefunction alone, to presume that such un-collapsing or re-merging is any less possible than the more familiar bifurcation or collapsing—in fact, the former is simply the time-reversal of the latter, and since quantum mechanics in itself is time-symmetric, they are essentially the same process. Of course, due to the thermodynamic arrow of time, real events in the real world do not experience this, as it would be stupendously unlikely (but, still, not impossible).

2.5.2 Entanglement and Nonlocality

We saw in Ch. 1 that a "superposition" is never a superposition of states of a subsystem, but must, in general, be considered only as a superposition of states of the entire system. The subsystems, in

fact, do not even have precise states. A consequence of this fact is the appearance of two related phenomena: "entanglement" and "nonlocality". These will be described below in terms of only two subsystems, but everything generalizes straightforwardly to three or more subsystems.

Definition 2.19. States $|\psi_A\rangle$, $|\psi_B\rangle$ of two subsystems represented in Hilbert spaces $\mathcal{H}_A$, $\mathcal{H}_B$ are "uncorrelated" or "unentangled" if $|\psi_{AB}\rangle$, the state of the composite system is "separable", meaning it can be expressed as a tensor product of the combined tensor product space, $\mathcal{H}_A \otimes \mathcal{H}_B$

$$|\psi_{AB}\rangle = |\psi_A\rangle \otimes |\psi_B\rangle \tag{2.145}$$

Whereas, if the state of the composite system *cannot* be so expressed, it is "inseparable", and the subsystems are said to be "correlated" or "entangled".

Entangled subsystems are expressed in the form of a superposition of the composite system.

Definition 2.20. If $\{|a_k\rangle\}$ and $\{|b_k\rangle\}$ are bases for A and B respectively, so that $|\psi_A\rangle = \sum_k c_k^A |a_k\rangle$ and $|\psi_B\rangle = \sum_k c_k^B |b_k\rangle$, then an "inseparable state" is expressed in the for

$$|\psi_{AB}\rangle = \sum_{i,j} c_{ij} |a_i\rangle \otimes |b_j\rangle \tag{2.146}$$

where $c_{ij} \neq c_i^A c_j^B$, since otherwise the above would not be a true superposition, since it could be straightforwardly re-expressed as $\sum_k c_k^A |a_k\rangle \otimes \sum_k c_k^B |b_k\rangle = |\psi_A\rangle \otimes |\psi_B\rangle$.

Theorem 2.21. ***The Schmidt theorem:*** *for bipartite state $|\psi_{AB}\rangle = |\psi_A\rangle \otimes |\psi_B\rangle$ in a tensor product Hilbert space $\mathcal{H}_A \otimes \mathcal{H}_B$, there will always be bases $\{|\alpha_k\rangle\}$ and $\{|\beta_k\rangle\}$ for $\mathcal{H}_A$ and $\mathcal{H}_B$ (each of dimensionality equal to the minimum of that of $\mathcal{H}_A$ and $\mathcal{H}_B$) such that the state can be written in the diagonal form of the "Schmidt decomposition"* [192]:

$$|\psi_{AB}\rangle = \sum_k c_k |\alpha_k\rangle \otimes |\beta_k\rangle \tag{2.147}$$

where the "Schmidt coefficients" $\{c_k\}$ are unique for the given state, and can be chosen to be non-negative and real, so that they yield a probability distribution $\{p_k\}$:

$$\{c_k\} = \{\sqrt{p_k}\} \tag{2.148}$$

This decomposition is unique in the case of nondegeneracy. There is no analogous guarantee for the existence of a diagonal decomposition of a tripartite state $|\psi_{ABC}\rangle = |\psi_A\rangle \otimes |\psi_B\rangle \otimes |\psi_C\rangle$. However, it can be shown that *if* there is such a tripartite decomposition, then it is unique [31, 29, 30] (in which case, we can simply combine two of the Hilbert subspaces $\mathcal{H}_B$ and $\mathcal{H}_C$ into one subspace $\mathcal{H}_{BC}$ to obtain the Schmidt decomposition).

Definition 2.22. One way of defining separability is that a state represented in such a tensor product Hilbert space is "separable" if (and only if) its Schmidt decomposition has one (and only one) non-zero Schmidt coefficient, else the state is "entangled".

Definition 2.23. If all the non-zero Schmidt coefficients are equal, then the pure state is "maximally entangled".

Entangled subsystems share mutual information, whereas unentangled systems are informationally isolated (neither has any information about the other). Of course, whether or not there is entanglement is a relative question, not a feature of the overall system itself, since it depends on how the overall system is separated (factored) into subsystems.

The classic example of entanglement are the four "Bell states", which are "maximally entangled" (so that each subsystem contains complete information about the other):

$$\begin{aligned}
&\frac{1}{\sqrt{2}}\left(|0\rangle_A \otimes |0\rangle_B + |1\rangle_A \otimes |1\rangle_B\right) \\
&\frac{1}{\sqrt{2}}\left(|0\rangle_A \otimes |0\rangle_B - |1\rangle_A \otimes |1\rangle_B\right) \\
&\frac{1}{\sqrt{2}}\left(|0\rangle_A \otimes |1\rangle_B + |1\rangle_A \otimes |0\rangle_B\right) \\
&\frac{1}{\sqrt{2}}\left(|0\rangle_A \otimes |1\rangle_B - |1\rangle_A \otimes |0\rangle_B\right)
\end{aligned} \tag{2.149}$$

It is also possible to have partially (non-maximally) entangled states, where each subsystem contains mutual (but incomplete) information about the other.

Entanglement of subsystems is "nonlocal", meaning that, since the superposition is of the entire composite system as a whole, that any "collapse" of this superposition (whether actual or perspectival) will appear as a discontinuous and nonlocal change. This is best introduced with the classic thought experiment put forward by Einstein, Podolsky and Rosen [74].

Consider the last of the Bell states above

$$\frac{1}{\sqrt{2}}\left(|up\rangle_A \otimes |down\rangle_B - |down\rangle_A \otimes |up\rangle_B\right) \tag{2.150}$$

to represent two spin-1/2 particles emitted from the same device, traveling in opposite directions, one with spin *up*, and one with spin *down*. Perhaps the particles are an electron/positron pair. They are entangled from the start, and remain entangled no matter how separated they become.

Now imagine that two observers, Alice and Bob, are situated at great distance from each other (assume they are light-years apart) in direct line to receive particle A and B respectively.

So long as coherence is maintained—so that Alice and Bob receive no information to distinguish between the two possible spins they might be about to observe—then the entire system is still in

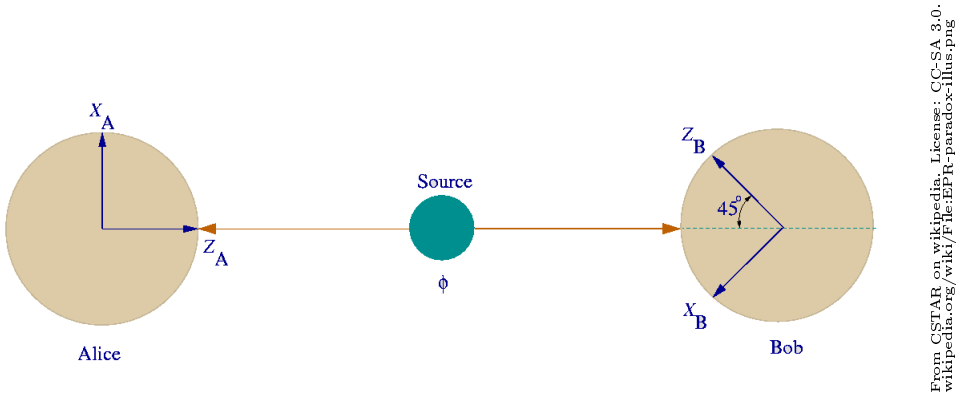


Figure 2.9: EPR entanglement

superposition. It is crucial to realize, as has already been stressed, that what we have here is *not* an electron in an *up*/*down* superposition travelling in one direction, and a positron in an *up*/*down* superposition, travelling in the opposite direction. Rather, the *entire system* is in a superposition of an *up* electron and *down* positron superposed with a *down* electron and an *up* positron, travelling in opposite directions. The two subsystems are inseparable, all the superposed states of one sharing total mutual information with its relative state in the other. However, Alice and Bob constitute two *more* subsystems, each of which is *separable* from the particle it is about to observe (putting aside the question of whether of not Alice and Bob are already entangled with each other). However, when Alice observes particle A (assume for now that Alice is a few light-seconds closer to the source than Bob), she now has total information about it, and she herself is now entangled with *both* particles. In many worlds terminology, this just means there is now a superposition of Alice observing an *up* electron while a *down* positron hurls toward Bob superposed with Alice observing a *down* electron while an *up* positron hurls toward Bob. Neither "world" involves any "nonlocal" (faster-than-light) communication or causal effects. However, in the language of wavefunction collapse, Alice's electron collapses into one state (say *up*) and *not* the other. At the same moment, the positron collapses into the corresponding relative state (since it was the entire system state in superposition, not the subsystems). However, since Alice's observation causes a discontinuous collapse into one and *only one* alternative, this means that information about which state she observed was somehow transmitted nonlocally, instantaneously and faster-than-light. Bob now has no alternative but to observe *down*, just because of what Alice did light-years away.

This thought-experiment really only constitutes an example of non-local causation if one assumes some variation of mechanical collapse. If one adopts wavefunction realism, then there is no non-local causation at all. Neither is there, as some MWI critics have suggested, a nonlocal "splitting" mechanism (so that, while each world is still causally local, the splitting mechanism itself acts nonlocally);

this is a misleading picture because, quite simply, world splitting is not a causal mechanism, but simply a matter of perspective.

This is not a purely quantum mechanical situation. Many examples can be given of non-quantum "faster-than-light" effects that are clearly not causal, but perspectival, and no causal paradox need be entertained to explain them. A classic example is the spotlight on the moon. If you shine a flashlight (or more realistically, a laser) from the Earth onto the surface of the moon, you can make the spot of light on the moon travel from one end of the moon to the other with just a tiny wrist movement, causing the spot of light to zoom across the surface of the moon faster than light. If you doubt that the speed could actually be faster than light, consider that, in principle, you could substitute for the moon another body at any distance from Earth that you desire. The apparent faster than light travel of the spotlight is just a matter of perspective, and not a real faster-than-light motion. Consider that you could not use this effect to transmit information across the moon faster than light. While the "spot" may "travel" faster than light, no physical object or energy is travelling that fast. The photons—and any information they carry—are travelling from the Earth to the moon, at light speed. If the light spot was flicked back and forth across the moon in a Morse code pattern, a message could indeed be sent, but it would be travelling from the Earth to the moon, not from one side of the moon to the other.

Assuming wavefunction realism, the situation for Alice and Bob is no different. In each "world", each particle travels no faster than light. When Alice observes *up* she learns which branch she is now in—and the (now) other Alice learns the same for *down*. But there is no nonlocal, discontinuous "splitting" of the universe, transmitted from Alice light-years away. Unfortunately for collapse (and many other related) scenarios, because only one event occurs for Alice, even from the perspective of the entire composite system, one is forced to conclude that this collapse event somehow made itself instantaneously felt light-years away.

Hidden-variables theories try to maintain that there are unseen variables in the system that somehow, behind the scenes of quantum mechanics, causally determine which state Alice will observe. However, Bell's theorem [16] proved that any such hidden variables must necessarily involve nonlocal correlations. Hence, the most popular hidden variables theory today, that of de Broglie-Bohm [19], makes non-locality in the universe an integral feature. The universe, Bohm has said, is an "undivided whole".

While I have said that nonlocal causal effects are not required in an MWI conception of EPR, it may not be entirely fair—even without nonlocal "splitting"—to define the MWI as an entirely "local" theory, although this is sometimes claimed of it. Since it is the entire system that is in superposition—while the individual elements each act according to local causation—it follows that

the very notion of "locality" here is fundamentally an optional "subsystem" concept. Without the division into subsystems, there is no superposition, no entanglement, and no locality. And since such subdivision is optional—and can be performed in many mutually incompatible ways—one would like to view the composite system as the more objective and non-perspectival view of the system. In other words, there is really only the entire system—the wavefunction of the universe. It cannot fundamentally be broken down into localized parts, except as a matter of arbitrary perspective. Such analyses may entail "perspectival" causal non-locality (which being perspectival is not real, as such). However, this non-decomposable nature of the whole, whose "parts" are not real unto themselves, while not the same as Bohmian "wholeness", is certainly its own kind of thorough-going holism, which is still, in its own way, "non-local". This is still not a *causal* nonlocality, however, since the very notion of the action of mechanistic causes (A causes B) is a matter of separation into parts. However, on the level of the entire system, we still appear to have a very strong kind of "undivided whole"—we could call it "systemic nonlocality", rather than "causal nonlocality"—that is as thorough-going as that in any collapse or hidden-variables theory.

2.5.3 Macroscopic Superposition

It is sometimes tempting to think that quantum paradoxes are only relevant to very tiny, microscopic systems. The Schrödinger's Cat gedanken experiment [194] shows that entanglement cannot be contained to the microscopic level, at least not in principle. If there is microscopic entanglement, unitary evolution implies macroscopic entanglement and macroscopic superpositions.

As in §1.2.4, we assume that a sample of radioactive material is in a box completely sealed from the environment, along with a cat and vial of hydrocyanic acid hooked up to the radioactive sample through a diabolical device. The radioactive sample has a 50% chance of decaying in the next hour, breaking the vial and killing the cat.

We start with the initial overall system state $|\psi\rangle$, which is a vector in Hilbert space $\mathcal{H}$. We will (arbitrarily) factor this system into two subsystems, "box" and "observer":

$$|\psi\rangle = |\psi_{box}\rangle\,|\psi_{observer}\rangle \tag{2.151}$$

Recall that $|\psi_{box}\rangle\,|\psi_{observer}\rangle$ is short-hand for the tensor product $|\psi_{box}\rangle \otimes |\psi_{observer}\rangle$, so this state is a vector in the tensor-product Hilbert space that combines the Hilbert space for $|\psi_{box}\rangle$ and the Hilbert space for $|\psi_{box}\rangle$:

$$\mathcal{H} = \mathcal{H}_{box} \otimes \mathcal{H}_{observer} \tag{2.152}$$

The "box" can be further factored into "vial" and "cat",

$$|\psi_{box}\rangle = |\psi_{vial}\rangle\,|\psi_{cat}\rangle$$
$$\mathcal{H}_{box} = \mathcal{H}_{vial} \otimes \mathcal{H}_{cat} \tag{2.153}$$

so that

$$|\psi\rangle = |\psi_{vial}\rangle\,|\psi_{cat}\rangle\,|\psi_{observer}\rangle$$
$$\mathcal{H} = \mathcal{H}_{vial} \otimes \mathcal{H}_{cat} \otimes \mathcal{H}_{observer}. \tag{2.154}$$

Other factorings are possible, but this will suffice for our purposes.

Before the clock starts on the one-hour wait, everything is in the initial state of $|\psi\rangle$. After the one hour has elapsed, just after the decay has happened (or not), and just before the decay (or lack thereof) has had a chance to affect anything else in the box, the overall system has evolved into state $|\psi'\rangle$, a superposition including both the relative decay state of the vial, $|\rightsquigarrow\rangle$, and the relative no-decay state, $|\nrightarrow\rangle$. The cat and observer will be in states of readiness, $|\text{cat}_{ready}\rangle$ and $|\text{observer}_{ready}\rangle$ respectively, as yet unaffected by any change in the vial:

$$|\psi\rangle \Longrightarrow |\psi'\rangle = \frac{1}{\sqrt{2}}(|\rightsquigarrow\rangle + |\nrightarrow\rangle)\,|\text{cat}_{ready}\rangle\,|\text{observer}_{ready}\rangle \tag{2.155}$$

None of the three subsystems have as yet become correlated with each other, since they can be written as a single tensor product. In the equation above, all three subsystems are separated. Recall that Schrödinger's contention was that it is in general the *entire system* that is in superposition, not individual subsystems. While this is true in general, it is still possible to have a subsystem—as above for the vial—that *is* thoroughly separated from its environment. However, in this case, since the "superposition" is isolated within a single subsystem, it is a pure quantum state, which could just as well be expressed as a single term, in some other basis. If this kind of superposition was all there was to quantum superpositions, there would be no measurement problem.

Note that, even though the three subsystems are separable here, it is still possible to write the above equation in the more general form, *as if* the subsystems were correlated:

$$|\psi\rangle \Longrightarrow |\psi'\rangle = \frac{1}{\sqrt{2}}\,|\rightsquigarrow\rangle\,|\text{cat}_{ready}\rangle\,|\text{observer}_{ready}\rangle + \frac{1}{\sqrt{2}}\,|\nrightarrow\rangle\,|\text{cat}_{ready}\rangle\,|\text{observer}_{ready}\rangle) \tag{2.156}$$

We can still see that the subsystems are not *really* correlated, however, since all but one of the subsystems are identical in both terms of the "superposition", and so can be factored out, cleaning separating the three systems, as in our previous notation.

Once the decay has had a chance to affect the cat subsystem (the cat system receives information about the vial system), the cat and vial become correlated or entangled, and the entire system is in

now in a superposition of containing a dead cat (in state $|\text{cat}_{dead}\rangle$) and a live cat (in state $|\text{cat}_{alive}\rangle$).

$$|\psi'\rangle \Longrightarrow |\psi''\rangle = \frac{1}{\sqrt{2}} |\rightsquigarrow\rangle |\text{cat}_{dead}\rangle |\text{observer}_{ready}\rangle + \frac{1}{\sqrt{2}} |\nrightarrow\rangle |\text{cat}_{alive}\rangle |\text{observer}_{ready}\rangle) \tag{2.157}$$

We can now only factor out the observer:

$$|\psi'\rangle \Longrightarrow |\psi''\rangle = \frac{1}{\sqrt{2}} (|\rightsquigarrow\rangle |\text{cat}_{dead}\rangle + |\nrightarrow\rangle |\text{cat}_{alive}\rangle) |\text{observer}_{ready}\rangle \tag{2.158}$$

Note that, while the cat no longer has a state of its own, so long as we restrict our attention just to the box subsystem, there is no entanglement. The box is in a "superposition" of "containing a dead cat" (state $|\text{box}_{dead-cat}\rangle$) and "containing a live cat" (state $|\text{box}_{live-cat}\rangle$).

$$|\psi'\rangle \Longrightarrow |\psi''\rangle = \frac{1}{\sqrt{2}} (|\text{box}_{dead-cat}\rangle + |\text{box}_{live-cat}\rangle) |\text{observer}_{ready}\rangle \tag{2.159}$$

Again, the idea of entanglement is all about subsystems, and is not an inherent feature of the system itself.

The terms of the thought experiment stipulate that the box allows no information or physical effects to flow outside of its walls until the box is opened. This is unrealistic, since it denies the possibility that the fate of the cat may affect the observer even before he opens the box and looks inside (so that the observer unwittingly receives information about the relative state of the cat without explicitly "looking"). However, these are the terms of the thought experiment, so we accept them for the sake of the argument—and there is nothing wrong with that, so long as we don't start thinking that the thought experiment proves the existence of half-dead-half-live cats in the real world.

Since the observer receives no information about the cat, so long as the lid of the box stays closed, he can remain uncorrelated with the inside of the box indefinitely. Once the box is opened, however, the individual observer receives information about the relative state of the cat, and becomes correlated with it. The observer system is thus now in superposition of "seeing-a-dead-cat" (state $|\text{observer}_{dead-cat}\rangle$) and "seeing-a-live-cat" (state $|\text{observer}_{ready}\rangle$):

$$|\psi''\rangle \Longrightarrow |\psi'''\rangle = \frac{1}{\sqrt{2}} |\rightsquigarrow\rangle |\text{cat}_{dead}\rangle |\text{observer}_{dead-cat}\rangle + \frac{1}{\sqrt{2}} |\nrightarrow\rangle |\text{cat}_{alive}\rangle |\text{observer}_{live-cat}\rangle) \tag{2.160}$$

None of the subsystem states can be factored out; all three subsystems are now entangled.

This thought experiment was intended to show that superpositions, taken literally, will expand to the macroscopic scale, and that conceptualizing them as a "smearing out" of actual states of things in the world is therefore not tenable. This highlights one of the key aspects of the measurement problem, sometimes called the "problem of definite outcomes". If such a superposition, given the unitary evolution of the wavefunction, expands to the macroscopic scale, then there is no definite

outcome to the experiment. How, then, is it that we actually do experience a definite outcome when we look inside the box? There seems to be a fundamental tension here, forcing us to choose between macroscopic superpositions (logically mandated by the quantum equations) and definite outcomes (empirically mandated by experience).

While Schrödinger's cat experiment was intended to show the untenability of the former position, as discussed in §1.2.4, Everett believed he had solved the measurement problem by accepting the reality of macroscopic superpositions, and rejecting (non-relative) definite outcomes as an empirical illusion of perspective.

Assuming it were possible to maintain the coherence of the box before the lid is opened (if the box were truly an absolute black box), then we can treat the box and the experimenter as separable, and can effectively write the state of the box as if it were a separate system:

$$\frac{1}{\sqrt{2}}\left(|\rightsquigarrow\rangle \otimes |\mathrm{cat}_{dead}\rangle + |\nrightarrow\rangle \otimes |\mathrm{cat}_{alive}\rangle\right) \otimes observer_{ready} \Longrightarrow$$
$$\frac{1}{\sqrt{2}}\left(|\rightsquigarrow\rangle \otimes |\mathrm{cat}_{dead}\rangle \otimes observer_{dead-cat} + |\nrightarrow\rangle \otimes |\mathrm{cat}_{alive}\rangle \otimes observer_{live-cat}\right) \quad (2.161)$$

Now when we, the experimenters, open the box to look inside, the standard interpretations employ the "collapse postulate" (or something similar) which requires that, because a measurement has now been made, there must be a collapse of the superposition into one of the two possibilities, and either we have

$$\frac{1}{\sqrt{2}}\left(|\rightsquigarrow\rangle \otimes |\mathrm{cat}_{dead}\rangle \otimes observer_{dead-cat} + |\nrightarrow\rangle \otimes |\mathrm{cat}_{alive}\rangle \otimes observer_{live-cat}\right) \rightarrowtail$$
$$|\rightsquigarrow\rangle\, |\mathrm{cat}_{dead}\rangle\, |\mathrm{observer}_{dead-cat}\rangle \quad (2.162)$$

or we have

$$\frac{1}{\sqrt{2}}\left(|\rightsquigarrow\rangle \otimes |\mathrm{cat}_{dead}\rangle \otimes observer_{dead-cat} + |\nrightarrow\rangle \otimes |\mathrm{cat}_{alive}\rangle \otimes observer_{live-cat}\right) \rightarrowtail$$
$$|\rightsquigarrow\rangle\, |\mathrm{cat}_{alive}\rangle\, |\mathrm{observer}_{live-cat}\rangle \quad (2.163)$$

Of course, in the MWI, we can still use this notation to represent the perspectival "collapse" of the wavefunction, but it does not represent a real physical process.

2.5.4 Environment-induced Decoherence

Decoherence theory [108, 236] allows us to actually quantify the degree of entanglement between two subsystems. From an information-theoretic point of view, the degree of entanglement is a matter of how much of subsystem A is actually encoded into subsystem B. The primary mathematical tool is the *reduced density matrix* (my explanation more or less follows that in [189]).

Definition 2.24. For a pure quantum state $|\psi\rangle$ ("pure" being the only kind we have so far considered), the "density matrix" is just the projector for that state:

$$[\psi] = |\psi\rangle \langle\psi| \tag{2.164}$$

We can tell if any given state is pure, by checking that

$$[\psi]^2 = [\psi] \tag{2.165}$$

In a particular basis:

$$\begin{aligned} |\psi\rangle &= \sum_k c_k |k\rangle \\ [\psi] = |\psi\rangle \langle\psi| &= \sum_{ij} c_i c_j^* |i\rangle \langle j| \end{aligned} \tag{2.166}$$

Quantum coherence between the $|i\rangle$ and $|j\rangle$ components are represented by the terms $c_i c_j^* |i\rangle \langle j|$ in this sum, which for $i \neq j$ are the "interference terms" or "off-diagonal terms" of $[\psi]$ in this basis, since they represent the off-diagonal terms in the operator matrix (the terms with a phase factor). There is always a basis in which $[\psi]$ is diagonal, and there are *no* off-diagonal terms.

Assuming $\hat{O}$ is Hermitian, so it can represent a physical observable, we now compute the trace of this operator projected onto our quantum state $|\psi\rangle$, using our Hermitian operator's diagonal basis, $\{|k\rangle\}$:

$$\begin{aligned} \mathrm{Tr}([\psi]\,\hat{O}) &= \sum_k \langle k|\psi\rangle\langle\psi|\hat{O}|k\rangle \\ &= \sum o_k |\langle o_k|\psi\rangle|^2 \\ &= \sum_k o_k p(k) \end{aligned} \tag{2.167}$$

So $Tr([\psi]\,\hat{O})$ is a weighted average of outcomes of $\hat{O}$, weighted by their Born probabilities; in other words, an expectation value:

$$\mathrm{Tr}([\psi]\,\hat{O}) = \langle\hat{O}\rangle \tag{2.168}$$

So far, we have been dealing with regular (or "pure") quantum states, for which this works out the same as the conventional calculation of expectation value:

$$\langle\hat{O}\rangle = \mathrm{Tr}([\psi]\,\hat{O}) = \langle\psi|\hat{O}|\psi\rangle \tag{2.169}$$

The real use for density matrices, however, is for impure or "mixed" states.

Definition 2.25. A "pure" quantum state is a state represented by a unit vector in Hilbert space. A "mixed" quantum state is simply a classical (*not* quantum) probability distribution over an ensemble of pure states.

Definition 2.26. The density matrix $[\psi]$ for a mixed state ψ is just a linear combination of the density matrices for the members of the (classical) ensemble, weighted by their (classical) probabilities:

$$[\psi] = \sum_k p_k \, [\psi_k] \tag{2.170}$$

where p_k is simply a classical probability measure, as distinguished from $p()$, the quantum probability. Note that we generalize here the [] notation used for projectors for mixed states, so that (unlike with a projector) the ψ inside $[\psi]$ does not here correspond to a pure state $|\psi\rangle$, and does *not* in general represent a projector $|\psi\rangle \langle\psi|$. It happens to be a simple projector, of course, in the special case of a pure state, but otherwise it is a probabilistically weighted sum of projectors. This weighted sum yields a classical-probabilistic version of the expected value calculation:

$$\left\langle \hat{O} \right\rangle = \sum_k p_k \left\langle \psi_k | \hat{O} | \psi_k \right\rangle \tag{2.171}$$

And since the density matrix rolls classical probabilities right into the pre-existing Born probabilities for the pure states, this expectation value can be computed the same way, whether the state is pure or mixed, using the trace:

$$\langle \hat{O} \rangle = \mathrm{Tr}([\psi] \, \hat{O}) \tag{2.172}$$

In other words, mixed-state density matrices layer classical probabilities on top of the regular quantum formalism. They are used when there is (classical) uncertainty as to which pure quantum state we are dealing with. Recall that a trace of $[\psi] \, \hat{O}$ already has the effect of probabilistically weighting an operator sum according to the (quantum or Born) probabilities. Now we weight each component of the mixed density matrix with the classical probabilities. We are doing a (classical) probability sum of (quantum) probability sums. The result functions just like a pure state density matrix; it just has the classical probabilities rolled into it. Note that the mixed-state summation has no off-diagonal terms, so involves no entanglement; it is purely classical in its properties.

The criteria for quantum entanglement or correlation between subsystems can be restated in terms of density matrices.

Definition 2.27. Subsystems represented in subspaces $\mathcal{H}_A$,$\mathcal{H}_B$,... of overall Hilbert space $\mathcal{H}$ are "correlated" if and only if it is *not* possible to factor the density matrix for the overall system into a tensor product of density matrices for the subsystems:

$$[\psi] = [\psi_A] \otimes [\psi_B] \otimes \cdots \tag{2.173}$$

The form of the density matrix most useful for analysis of environmental decoherence is the "reduced density matrix". Unlike with the mixed-state density matrix, here we do not (necessarily)

have any classical uncertainty about which state we have. It is used rather for composite systems, where we have factored the system into two (or more) subsystems, such as A and B, one of which (say A) is typically a measuring apparatus or somesuch. Reduced density matrices are especially useful when it is not possible (or at least not practical) to make any measurements on one of the subsystems (usually, this inaccessible subsystem is the "environment").

Assume, then, that we have a composite system in pure state $|\psi\rangle$, represented in tensor product space $\mathcal{H}_A \otimes \mathcal{H}_E$, with orthonormal bases $\{|k\rangle\}$ and $\{|k'\rangle\}$ for subspaces $\mathcal{H}_A$ and $\mathcal{H}_E$ respectively:

$$|\psi\rangle = \frac{1}{\sqrt{2}} \left(|a_1\rangle \otimes |e_1\rangle + |a_2\rangle \otimes |e_2\rangle \right) \tag{2.174}$$

yielding the (pure state) density matrix

$$[\psi] = \frac{1}{2} \sum_{i,j=1}^{2} |a_i\rangle \langle a_j| \otimes |e_i\rangle \langle e_j| \tag{2.175}$$

A "reduced density matrix" is obtained by use of the "partial trace" of A over E, $[A/_E]$, which is defined as $Tr_E([\psi])$, the trace of ψ using an orthonormal basis $\{|e_k\rangle\}$ for E:

$$[A/_E] = \mathrm{Tr}_E([\psi]) = \sum_k \langle e_k \, [\psi] \, e_k \rangle \tag{2.176}$$

Now instead of using $\hat{O}$, we use $\hat{O}_A \otimes \hat{I}_E$, since we cannot measure E. We can carry out the E part of the trace (since we are just using identity), and use an orthonormal basis for E in calculating the trace for A. In other words, we trace the density matrix for A over the degrees of freedom of E:

$$\begin{aligned} [A/_E] &= \sum_k \langle e_k \, [\psi] \, e_k \rangle \\ &= \frac{1}{N} \sum_{i,j}^{N} |a_i\rangle\langle a_j| \langle e_j | e_i \rangle \end{aligned} \tag{2.177}$$

where N is the number of components of $|\psi\rangle$. Note that the off-diagonal "interference" terms are essentially weighted (in a kind of averaging) by the degree of overlap between the correlated environmental states. When the correlated environmental states have perfect distinguishability, $\langle e_1 | e_2 \rangle = 0$, then the reduced density matrix is diagonal in the $\{|a_1\rangle, |a_2\rangle\}$ basis, and there are no interference terms. Thus, complete information is encoded in the environment about the apparatus, and the subsystems are maximally entangled. When there is large overlap, on the other hand, there is large mutual information between the subsystems, and they encode information about each other, even if the environment is effectively inaccessible to the observer.

Computing the expectation values (see [189, p.48]), we get :

$$\begin{aligned}\langle \hat{O} \rangle &= \mathrm{Tr}([\psi]\,\hat{O}) && (2.178)\\ &= \mathrm{Tr}_A([A/_E]\,\hat{O}_A) && (2.179)\\ &= \mathrm{Tr}_A(\mathrm{Tr}_E([\psi])\hat{O}_A) && (2.180)\end{aligned}$$

For example, returning to the Schrödinger's cat experiment, we have the system $B \otimes E$ (B for "box" and E for environment). Measurements are made only on the box. We use $|d\rangle$ for "box with dead cat", and $|a\rangle$ for "box with live cat", for the states of the B. We use $|e_d\rangle$ and $|e_a\rangle$ for corresponding states in the environment that correlate with the $|d\rangle$ and $|a\rangle$ states, respectively. So these might correspond to the exact angle at which a photon scatters off the outside of the box, at a slightly different angle $|e_d\rangle$ when then cat is dead than the angle $|e_a\rangle$ it takes when the cat is alive.

We get the following reduced density matrix for B traced over E:

$$\begin{aligned}[B/_E] &= \frac{1}{2}\big(|d\rangle\langle d|\langle e_d|e_d\rangle + |a\rangle\langle a|\langle e_a|e_a\rangle + |d\rangle\langle a|\langle e_a|e_d\rangle + |a\rangle\langle d|\langle e_d|e_a\rangle\big)\\ &= \frac{1}{2}\big(|d\rangle\langle d| + |a\rangle\langle a| + |d\rangle\langle a|\langle e_a|e_d\rangle + |a\rangle\langle d|\langle e_d|e_a\rangle\big) \qquad (2.181)\end{aligned}$$

The first two terms, $\frac{1}{2}|d\rangle\langle d| + \frac{1}{2}|a\rangle\langle a|$, are the diagonal elements of the density matrix for the box. The remaining terms, $\frac{1}{2}|d\rangle\langle a|\langle e_a|e_d\rangle + \frac{1}{2}|a\rangle\langle d|\langle e_d|e_a\rangle$, are the off-diagonals, and these term represent the entanglement with the environment. Entanglement goes to zero as the bra-ket coefficients go to zero, which occurs when the environmental states are (approximately) orthogonal to one another, and the reduced state of the box is approximately

$$[B/_E] \approx \frac{1}{2}\big(|d\rangle\langle d| + |a\rangle\langle a|\big) \qquad (2.182)$$

In this case, these states are highly distinguishable, one from the other, and hence they are able to distinguish between the states of the box that they correlate to. When they are not orthogonal, they have a high degree of overlap, and are difficult to distinguish, and hence they are less able to distinguish—encode information about—the correlated states in the box. In the extreme case, the two states are equal, and there is no ability to encode information at all.

The interference terms have thus effectively disappeared. Note, however, that $[B/_E]$ is the *reduced* state of the *box*, wherein entanglement with the environment is considered inaccessible. The entire box-environment system, $B \otimes E$, remains an entangled (pure) state,

$$\begin{aligned}|B \otimes E\rangle &= \frac{1}{\sqrt{2}}\left(|d\rangle \otimes |e_d\rangle + |a\rangle \otimes |e_a\rangle\right) && (2.183)\\ [B \otimes E] &= \frac{1}{2}\left(|d\rangle\langle d| \otimes |e_d\rangle\langle e_d| + |a\rangle\langle a| \otimes |e_a\rangle\langle e_a|\right) && (2.184)\end{aligned}$$

so that entanglement with the environment is still very real; we just ignore it when we assume it is inaccessible.

When a measurement is made on the box, understood now in terms of $[B/_E]$, we still use something very much like the standard projective measurements, except that now we have these correlations with a larger box-environment system, of which the box is only a part. So our "box" is no longer a pure state prior to measurement, as it is in the classic Cat experiment.

The preferred states will be the eigenstates of the interaction Hamiltonian, $\hat{H}_{\text{int}}$, the component of the system Hamiltonian responsible for system-environment interaction. The overall system Hamiltonian, $\hat{H}$, can be divided into three components:

$$\hat{H} = \hat{H}_{\text{sys}} + \hat{H}_{\text{env}} + \hat{H}_{\text{int}} \tag{2.185}$$

for a factoring of the overall system into two Hilbert subspaces. This can be generalized to factoring into m subspaces, in which case, we have $\frac{1}{2}\left(m^2 + m\right)$ Hamiltonians: m subsystem Hamiltonians, and $\frac{1}{2}\left(m^2 - m\right)$ interaction Hamiltonians.[19] Of particular interest in measurement theory, especially when using synthetic *a priori* methods, is the case of three specific subsystems[211]: an *observer* (obr), an *observable* (obl) and an *environment* (env):

$$\hat{H} = \hat{H}_{\text{obr}} + \hat{H}_{\text{obl}} + \hat{H}_{\text{env}} + \hat{H}_{\text{obr-obl}} + \hat{H}_{\text{obr-env}} + \hat{H}_{\text{obl-env}} \tag{2.186}$$

In this case, we can also still speak of an interaction Hamiltonian, $\hat{H}_{\text{int}}$, for the total system:

$$\hat{H}_{\text{int}} = \hat{H}_{\text{obr-obl}} + \hat{H}_{\text{obr-env}} + \hat{H}_{\text{obl-env}} \tag{2.187}$$

In real-world systems with large, complex environments, $\hat{H}_{\text{int}}$ is responsible for most of the evolution of the system (its energy scales will be much greater than those of the Hamiltonians of individual subsystems). So we will find stable states under this Hamiltonian alone, the preferred states that commute with the interaction Hamiltonian. Given that the elements of $B = \{|b_1\rangle, \ldots, |b_n\rangle\}$ are the eigenstates of $\hat{H}_{int}$ pertaining to the system being measured, the projectors of these states can be linearly combined to provide the preferred operators $\left\{\hat{O}_{b,o}\right\}$ for the system:

$$\hat{O}_{b,o} = \sum_{k \in b} o_k |b_k\rangle\langle b_k| \tag{2.188}$$

where $b \subseteq B$ and $o = \{o_k\}$ is an equal-sized set of arbitrary coefficients.

Since the $|b_k\rangle$ are eigenstates of $\hat{H}_{\text{int}}$, their projectors and the $\hat{O}_{b,o}$ commute with $\hat{H}_{\text{int}}$.

Keep in mind that the application of reduced density matrices usually assumes the Born rule, and so any *a priori* Born rule proof cannot rely on the usual reduced density matrix formulation of decoherence, although this does not mean it cannot appeal to decoherence (see, for example, [238]).

[19] There are $1/2\left(m^2 - m\right)$ interaction Hamiltonians because we need one for each of the m^2 pairs of subsystems, $(\mathcal{H}_i, H_j)$, excepting the m diagonal cases $(\mathcal{H}_k.\mathcal{H}_k)$, and the redundant half of what remains, since $(\mathcal{H}_i, \mathcal{H}_j) = (\mathcal{H}_j, \mathcal{H}_i)$.

2.5.5 POVMs

Assume we are measuring the "nearly-orthogonal" states $|d\rangle$ and $|a\rangle$ (for "dead" and "alive" cat states, perhaps). In the projective view, these form a preferred orthonormal basis, which can be used as the basis for our projective measurement operator. However, we can see from the full reduced density matrix, that in all likelihood, these states will only be "nearly" orthogonal. There will be some small amplitude in the off-diagonals. Hence, a more general model of measurement would allow (potentially) non-orthogonal measurement operators.

A "positive operator valued measure" or "POVM" is a set of Hermitian, positive semidefinite operators,

$$\{\lceil k \rceil : k = 1 \cdots N\} \tag{2.189}$$

that constitute a decomposition of identity,

$$\sum_k \lceil k \rceil = \hat{I} \tag{2.190}$$

I use the "unfinished projector" notation $\lceil\ \rceil$ to signal that these operators can be used and thought of much like projectors, the only difference being that they are not required to be mutually orthogonal (although they may well be *very nearly* orthogonal). Projective measurement, in fact, can simply be considered a special case of POVM, where the operators happen to be mutually orthogonal. The projective case may be thought of as a simplified, idealized case—much like other idealized models in physics, such as frictionless planes and the like.

Positive semidefiniteness means nonnegative eigenvalues, which represent the probabilities of the outcomes that correspond to each operator in the POVM, given the pre-measurement state $|\psi\rangle$:

$$p(k|\psi) = \langle \psi \lceil k \rceil \psi \rangle \geq 0$$

We call N the "dimension" of the POVM (which can be *larger* than the dimension of the Hilbert space, since the operators need not be orthogonal).

Theorem 2.28. *If, for all k, there is an $\hat{E}_k$ such that $\lceil k \rceil = \hat{E}_k^* \hat{E}_k$, then the $\{\lceil k \rceil\}$ are positive semidefinite and Hermitian with nonnegative eigenvalues $p(k) = \langle \psi \lceil k \rceil \psi \rangle$, such that*

$$\sum_k p(k) = 1 \tag{2.191}$$

The state of the system after measurement is:

$$\frac{1}{\sqrt{p(k|\psi)}} \hat{E}_k |\psi\rangle \tag{2.192}$$

Such a generalized measurement is actually just a projective measurement if formulated fully in terms of the larger system (Naimark's theorem [150, Th.4.6]). Bringing this all back to the cat in the box, when you open the box and look inside, you are performing (in the original experiment) a projective measurement. Before looking inside, the cat is literally in a superposition of alive and dead, which collapses/branches when you look inside, because you are performing a projective measurement that projects the box onto one of the axes of the dead/alive basis. The mathematics of projective measurement does not, of course, explain why looking inside the box did not project the cat onto an axis of some *other* basis, so that there were still components of deadness and aliveness in the cat's state, relative to your state. So we still have a preferred basis problem, although now we are dealing more correctly with a POVM rather than a basis.

Definition 2.29. A "preferred POVM" for a particular measurement situation is a POVM for which each operator corresponds to exactly one (synthetically) possible outcome of the measurement. The members of a preferred POVM are "preferred positive semidefinite operators" or "effects". A member of a POVM that is *not* necessarily an effect is a "potential effect".

The unqualified word "effect" in the literature more often means "potential effect" (*i.e.,* any member of a POVM), but (with our ASU context in mind) I reserve it here for those members of POVMs that are actually potentially observable in a manner that retains the synthetic unity of the observer. In this context, "effect" has a necessary synthetic aspect, and means more or less the same thing as the common English word "effect".

The existence of a preferred POVM is just a generalization of the idea of a preferred basis. From the relative state point of view, it is not really a mystery, once synthetic unity is factored in, that there will be some potential effects that correspond to true effects (*i.e.,* to distinct conscious states), and some that do not. Thus, some POVMs will be preferred simply in order to maintain the synthetic unity of consciousness of the observer. By this reasoning, the fundamental justification for there being a preferred basis/POVM is the unity of consciousness, *not* environmental decoherence. Environmental decoherence gives much insight into how such a preferred basis gets selected out, and how it can converge on something like classical behavior, but this does not strictly solve the preferred basis/POVM problem.

In the decoherence view of Schrödinger's Cat, you are performing a POVM measurement when you open the box, *not* a projective measurement, so you are only *approximately* projecting onto an axis of the dead/alive basis. This means that the actual dead/alive "near-projection" can occur long before you actually open the box or become explicitly conscious of the result, since the correlation between your conscious state and the result is mediated through correlation with an effectively inaccessible environment. Because the off-diagonals of the density matrix are nearly zero, your

measurement *appears* to be projective, for all intents and purposes. And, in fact, it actually *is* projective if we fill in all the details of the information stored in the environment, since a POVM measurement of a subsystem is really just a projective measurement on the whole system, where we leave out the information in part of the system (here, the environment), which can be considered, for all intents and purposes, to be "lost" to the observer (so that, to the observer, it is just heat).

The preferred POVM is exactly the one that *does* decay the off-diagonals, because *synthetic unity demands a basis that is as diagonalized as necessary to maintain unity of consciousness*, else we would have interference between the different mental states in the different branches, which would imply that they were *not* separate mental states after all, and it would not be reasonable to characterize them in terms of branching/collapse (although analytically, such a description of the measurement event would be perfectly allowed; it is only disallowed when speaking *relative* to your current mental state—*i.e.,* from your personal perspective).

The correlated environmental variables hence do not need to be anything dramatic like a pointer device being in a certain macroscopic position. In cases where tiny effects in the environment can affect (or encode information about) whether the cat is dead or alive, then the correlated environmental states will be nearly orthogonal to each other, and the system state will be *nearly* decoherent, even though the "which-path" information is stored in some tiny environmental variable, like the exact angle that a photon took when bouncing off the outside of the box. To determine that it is *not* so would require massive information to be analyzed about the environment, which is why the observer can consider such environmental information to be lost information (heat).

In other cases, perhaps it takes a large effect in the environment to record the information about the cat, in which case, maintaining effective coherence of the system is easier. But it turns out that this is *very hard* to accomplish for most systems of much complexity. Why this is so can be appreciated with an informal argument from nonlinear dynamics: systems with positive "Lyapunov exponents" have trajectories in phase space that diverge exponentially due to tiny differences in initial conditions (a phenomenon known as "chaos", and sometimes "deterministic randomness", since it amounts to effective randomness even when the system is strictly deterministic). So long as the correlated environmental state behaves chaotically, even a tiny difference in the environment can produce a major effect on the system. So the off-diagonal terms in the reduced density matrix will be effectively suppressed *very* quickly, even if there is no overt information path between the box and the relevant environmental variables. In fact, complex systems in the real world tend to be replete with chaotic elements, so finding appropriately chaotic variables is not generally difficult. Zurek has shown that just the scattering of photons off the outside of the cat's box will be more than enough to create effective decoherence in under one second [236]. In such cases, the off-diagonal elements

may actually decay at an exponential rate. This is called "environment-induced decoherence".

2.5.6 Von Neumann Entropy

The eigenvalues $\{p_k\}$ of the reduced density matrix $[\psi]$ for a quantum state $|\psi\rangle$ can be used to calculate the "von Neumann entropy" [223], S or $S(\psi)$:

$$\begin{aligned} S(\psi) &= -\sum_k p_k \log p_k \\ &= -\mathrm{Tr}([\psi] \log [\psi]) \end{aligned} \tag{2.193}$$

If the intended state or system is understood, we can just write S. This is a straightforward application of Shannon entropy [199] to the density matrix (assuming that its diagonals define a probability distribution). Shannon entropy is the information-theoretic basis for entropy/disorder/information in statistical thermodynamics, communications theory, as well as many other fields. It is formally equivalent to algorithmic information, in which form it will be the basis for our response to the Born rule objectors in Ch. 6-8.

For pure states, by convention, the entropy is 0, so only mixed states have non-zero entropy, which must be positive. This is merely a kind of normalization, not an absolute requirement, but under this assumption, the von Neumann entropy becomes a measure of the degree of impurity of $[\psi]$. This makes intuitive sense if we are using reduced density matrices, as it is only when our system is said to be part of a larger system, of which we lack complete knowledge, that it makes sense to talk about entropy, and hence things like "heat" and "information loss".

Definition 2.30. The Shannon entropy $\overline{H}(A)$ of an information source A is simply the average "marginal entropy" or "self-information", $H(k)$, of each individual element k of the source:

$$H(k) = -\log p_k \tag{2.194}$$

So if x has probability 1/8, it will take $\log_2 8 = 3$ bits of information to communicate x. If it takes more or less the same number of bits to communicate any of the elements, then the distribution is highly disordered, and entropy is at a maximum for that size N of sample space. Namely, for all k:

$$p_k = \frac{1}{N} \tag{2.195}$$

$$S = \overline{H}(k) = \log N \tag{2.196}$$

As we would expect of an information measure, von Neumann entropy is basis-independent and invariant under unitary transformations (which is why unitary transformations are said to be "information-preserving"). Since it measures information, it can be used to measure the degree

of entanglement between quantum systems [80]. If quantum states are unentangled, they will contain no shared, or "mutual" information. The mutual information between two parts is simply the sum of the entropies of the parts minus the entropy of the whole (the "joint entropy").

The "joint entropy" of quantum states $|\psi\rangle$ and $|\varphi\rangle$ is simply $S(\psi \otimes \varphi)$.

Definition 2.31. The "mutual information" of quantum states $|\psi\rangle$ and $|\varphi\rangle$ is therefore defined as

$$H(\psi : \varphi) = S(\psi) + S(\varphi) - S(\psi \otimes \varphi) \tag{2.197}$$

which is, again, a direct analog of mutual Shannon information.

Clearly, for unentangled states,

$$S(\psi : \varphi) = 0 \tag{2.198}$$

while for entangled states,

$$S(\psi : \varphi) > 0 \tag{2.199}$$

Since the von Neumann mutual entropy provides a measure of degree of entanglement, in the context of environmentally-induced decoherence, this means degree of entanglement with the environment.

2.5.7 Schumacher Compression

Schumacher's coding theorem [195] placed von Neumann entropy on a firmer information theoretic footing by proving it to be a measure of the optimal encoding for quantum states. This discovery later became one of the foundation stones of quantum information theory. The theorem is a straightforward generalization to the quantum case of Shannon's noiseless coding theorem, which proves that Shannon entropy is a measure of the optimal encoding for messages following a known probability distribution, or more precisely that encoding errors can be reduced to arbitrarily low probability using the Shannon entropy measure.

Theorem 2.32. *Shannon's Noiseless Coding Theorem.*

Let A be a message source with $\overline{H}(A)$ its Shannon entropy, then for any $\delta, \epsilon > 0$:

(i) If $\overline{H}(A) + \delta$ bits are available per A message, then for sufficiently large N, sequences from A of length N can be coded into binary sequences with probability of error less than ϵ.

(ii) If $\overline{H}(A) - \delta$ bits are available per A message, then for sufficiently large N, if sequences from A of length N are coded into binary sequences, the probability of error will be greater than $1 - \epsilon$.

In other words, the optimal compression of messages from A is $\bar{H}(A)$ up to an additive constant, or $\bar{H}(A) + O(1)$.

See Ch. 4 for more details on exactly how the entropy measure is used to create an encoding. The basic idea is simple: use fewer bits to represent messages that are more probable, and you will, on average, save bits. In practice, the true probabilities of individual messages will not be known, and are in fact uncomputable, but can be guessed and in many cases closely approximated.

Schumacher's theorem generalizes this result straightforwardly to the quantum case.

Definition 2.33. A "qubit" or "quantum bit" is the "quantum information" (meaning von Neumann entropy) of a quantum state $|\psi\rangle$ that is in an equal superposition of storing both the $|0\rangle$ and $|1\rangle$ 1-bit values (each of which on its own is classically a single binary bit of information). For instance, we could use

$$|\psi_{01}\rangle = \frac{1}{\sqrt{2}}|0\rangle + \frac{1}{\sqrt{2}}|1\rangle \tag{2.200}$$

to store a qubit. The qubit can be considered a straightforward quantum generalization of the classical notion of "bit" (see Ch. 4 for more technical details on the classical version). Schumacher's theorem is a straightforward generalization of Shannon's theorem (for coding classical messages in bits) to the quantum case (of coding quantum messages in qubits).

Theorem 2.34. ***Schumacher's Quantum Noiseless Coding Theorem.*** [195, p.9]

Let M be a quantum signal source with a signal ensemble described by the density operator $[\psi]$, with $S(\psi)$ its von Neumann entropy, then for any $\delta, \epsilon > 0$:

(i) If $S(\psi)+\delta$ qubits are available per M signal, then for sufficiently large N, groups of N signals from the signal source M can be transposed via the available qubits with fidelity $F > 1 - \epsilon$.

(ii) If $S(\psi) - \delta$ qubits are available per M signal, then for sufficiently large N, if groups of N signals from the signal source M are transposed via the available qubits, then fidelity $F < \epsilon$."

In other words, the optimal compression of a quantum state in a superposition described by $[\psi]$ is $S(\psi)$ up to an additive constant, or $S(\psi) + O(1)$. Thus, Schumacher proved that von Neumann entropy measures the number of bits in the optimal data compression of a quantum state.

In Ch. 6-8, I will argue for a view of quantum probability based on algorithmic information theory, where the information content, or entropy, of a quantum state is identified with the number of bits in the optimal data compression of that state. So long as we stick to unitary evolution, we will find that entropy does not change, and we might as well normalize to a baseline of zero entropy. But when there is wavefunction collapse (or observer branching), we then have multiple such states to choose from, and the entropy *differences* yield a probability distribution. And since information theory gives us a way to go directly from entropy to probability (see Ch. 4), it might seem that Schumacher's theorem is tantamount to a Born rule proof, at least within the context of an algorithmic (compression-based) view of probability. However, this is not the case. Schumacher

compression might possibly define the *goal* for an algorithmic Born rule proof, but it does not deliver the proof itself, since both von Neumann and Schumacher based their formulations on the standard interpretation of the density matrix, with the Born rule assumed as part of the package. Schumacher, in other words, only showed that von Neumann entropy could be used to optimally compress quantum states, *if* the Born rule is true. He did not prove anything at all about how to optimally compress a quantum state simply given a finite description of that state, without any guidelines for calculating associated prior probabilities. As such, von Neumann entropy is, in a way, just a straightforward expression of the Born rule in information theoretic terms.

3 The Relative State Interpretation

Everett's relative state interpretation [79, 80] is built on three fundamental philosophical postulates:

Assumption 3.1. *Wavefunction realism*: the wavefunction is all there is; any physical system can be understood entirely in terms of this entity, without postulating the existence or action of anything else. This includes observation processes, which "are to be described completely by the state function of the composite system which includes the observer and his object-system, and which at all times obeys the wave equation." [80, p 8]

Assumption 3.2. *Psychophysical parallelism*: this phrase has been used in various ways by different people, but Everett [80, p 7] means by it von Neumann's version, that mental processes are wholly explainable *as if* they were physical processes: "...it is a fundamental requirement of the scientific viewpoint—the so-called principle of the psycho-physical parallelism—that it must be possible so to describe the extra-physical process of the subjective perception as if it were in reality in the physical world—*i.e.*, to assign to its parts equivalent physical processes in the objective environment, in ordinary space." [223, p 418-419]

Assumption 3.3. *Servomechanism equivalence*: observers can be explained wholly *as if* they were mechanical automata, in particular mechanical feedback control systems, or servomechanisms. Everett presents this as if it were almost a necessary consequence of psychophysical parallelism, that we "conceive of mechanical devices..., obeying natural laws, which we would be willing to call observers... [that] can be conceived as automatically functioning machines (servomechanisms) possessing recording devices (memory) and which are capable of responding to their environment. The behavior of these observers shall always be treated within the framework of wave mechanics. " [80, p 7,9]

It is perhaps arguable that wavefunction realism is the only really fundamental postulate here, and that the other two follow necessarily from it—however, some may disagree with that assessment, and Everett does explicitly state all three postulates, ensuring that the meaning and consequences of wavefunction realism, as he sees them, are clear.

Psychophysical parallelism seems very similar to materialism or physicalism; however, there is no claim here that matter/energy or the "physical" are the ultimate metaphysical constituents of reality, so this is certainly not *metaphysical* materialism. It may perhaps be taken, for our purposes, as a kind of operational materialism with respect to mind: subjective perceptions can be explained *as if* they were physical (not that they ultimately *are* physical).

This distinction will be important later on, since—as we develop a more and more algorithmic view of the wavefunction—it may eventually become important to clarify whether it is the "physical" or the "algorithmic" that is more fundamental. In either case, it could be argued that psychophysical parallelism follows from wavefunction realism, which declares that the wavefunction–which is, after all, entirely algorithmic—constitutes a complete description of any physical system.[20]

The third postulate could also arguably be inferred from the second, and Everett, in fact, states that the third postulate is required if we are to "do justice" to the second postulate. Nonetheless, it is important—especially for us—to note that Everett is explicitly assuming that psychophysical parallelism permits treating an observer as if it were a mechanical robot, since this is central to his attempt to prove the Born rule.

There is a fourth potential postulate that one might further infer from Everett's third:

Assumption 3.4. *Strong AI Postulate*: an appropriately programmed digital computer would have a mind in the same way a human being has a mind (meaning that it would be conscious in the same way a human is conscious).

I mention strong AI because it is a widely talked about idea that most people would probably take to be roughly the same thing as servomechanism equivalence. After all, if we can take observers to be effectively robots, then how could we deny that a human mind is simply the running of software in a digital computer? However, the strong AI postulate is stronger than Everett's servomechanism postulate in at least a couple of ways. Firstly, Everett's assumption only assumes that we can take human observers to be machines for the purposes of talking about observation in quantum mechanics. There may be some who would agree with this postulate, but disagree that a non-organic machine could literally have a conscious mind. Secondly, Everett's assumption would, in principle, include both digital and analog computers, and there may be some who would view the analog/digital distinction as important. In this dissertation, however, none of these distinctions play a strong role, and the reader may feel free to read "strong AI" in place of "servomechanism equivalence", unless a distinction is explicitly drawn.

[20] However, by the same token, one could also argue that some kind of metaphysical computationalism also follows from this, and it is not clear that Everett would go that far—this is why it is useful to explicitly state all three of Everett's explicit postulates, since not everyone will necessarily agree that they all follow necessarily from the first.

Contrary to common belief, Everett himself never claims that the wavefunction formalism directly yields his (and only his) interpretation. Rather, he claims only that it does so *if* one accepts his three basic postulates. Once assent is given to this philosophical framework, Everett proceeds to deal with the measurement problem by

1. rejecting the collapse postulate (by wavefunction realism), and
2. taking the observer, apparatus and environment as all quantum objects (by psychophysical parallelism), and
3. taking all resulting, superposed outcomes of an observation as being equally "real" branches of the wavefunction.

There is some controversy as to whether "equally real" here is supposed to mean that the branches are real *qua* branches, or whether it simply means that each branch is as real, and in the same way (whatever way that is) as any other branch. I read Everett as saying the latter, but, as we will see later, the way he proceeds with his Born rule proof could be taken to imply something more like the former. As a result Everett's treatment is a bit ambiguous with respect to the "real" nature of the individual branches.

In either case, there is definitely no fact of the matter as to which element of a superposition is realized after a measurement. All are realized. The observer himself is in superposition—he has "split" into two or copies of himself, effectively in different "worlds".

In general, if we have a superposition state $|\psi\rangle$, expressible in terms of $\{|b_k\rangle\}$ set of orthonormal basis states:

$$|\psi\rangle = \sum_k |b_k\rangle \langle b_k| \psi\rangle \tag{3.1}$$

and observer $|o\rangle$ makes an observation corresponding to an Hermitian operator diagonalized by $\{|b_k\rangle\}$, then the result is the new state,

$$|\psi, o\rangle = \sum_k |b_k\rangle \langle b_k| \psi, o\rangle \Longrightarrow |\psi, b_r, o_r\rangle \tag{3.2}$$

where $|b_r\rangle$ is the outcome of the measurement "relative to" observer state $|o_r\rangle$, which is resulting state of the pre-measurement observer $|o\rangle$. It occurs with probability as specified by the Born rule:

$$p(r|\psi, o) = |\langle b_r| \psi\rangle|^2 \tag{3.3}$$

The idea here is that only the unitary wavefunction evolution is required to understand what happens in a measurement. This makes for a very simple theory, in which the wavefunction itself is all there is, with no extra *ad hoc* collapse or other additions needed (or so the MWI advocates claim). However, the problems of justifying *(1)* a preferred basis and *(2)* the Born rule will be used by critics, as we will see, to argue that the MWI is *not* based solely on the wavefunction alone.

3.1 Objections to Many Worlds

Numerous problems have been proposed with the relative state interpretation. While I plan to focus on the Born rule objection, I will provide some background and discussion of other major objections, since the issues dealt with in these objections are not unrelated to the issues we will be seeing in our discussion on the Born rule.

3.1.1 The Everything-Happens Objection

This objection [7, 144, 110] claims that there is a fundamental problem with saying that every possible outcome happens (as in the MWI), and then trying to make probability claims *of any kind at all*. Here, the Born rule itself is not particularly the issue, but the very notion that probability statistics can make sense in the first place in a system where absolutely everything happens. Neumaier [144], for instance, states that "statistics makes no sense in MWI as everything that can happen, happens".

Response In my opinion, this objection on its own holds little weight. The quantum probabilities in the MWI do not refer to the probability of something *absolutely* happening or not happening, as is made crystal clear in Everett's original paper. Such probabilities are not absolute, but always relative to a single mind or observer, the particular state of which is relative to one outcome in the superposition, and not the others. So there is no sense in which "everything happens" from the viewpoint of any particular observers. And there is no reason to ask for probabilities from the viewpoint of the wavefunction itself, which seems to be something like what is being asked for in this objection.

The synthetic-analytic distinction helps us keep things straight in this case. If we insist that the "*a priori*" truth we are trying to prove is analytic, we can only look for something from the point of view of the wavefunction itself. But if we allow ourselves minimal *a priori* synthetic assumptions, such as are required to even speak about observations or outcomes in the first place, then there is no reason to claim that "everything that can happen, happens."

To make more of this objection one would have to object to the way in which superposed outcomes are relativized to the perspective of different, superposed minds in Everett's scheme, which is exactly what the next objection does.

3.1.2 The Superposed Minds Objection

Objections in this next class [7, 15] seem to arise from feelings of discomfort with the very idea of multiple minds/observers or incompatible mental states being in superposition—Albert talks about

"our very deep conviction that mental states never superpose" [79].

Response Of course, "our" usually means "my", since this conviction, as deep as it may be for some, is by no means shared by the rest of us. I have, personally, no conception of what this conviction is, or may feel like, since I have never felt it. And, since it has no logical defence, but is based solely on the idea of a shared conviction, it can hold no water for anyone who is unlucky enough never to have experienced this particular feeling of conviction. It is thus not really subject to public debate or discourse.

Everett suggests that this deep conviction is very much like the conviction that the Earth doesn't move, because we do not feel it move. For someone who has grown up without the Galilean concept of relativity of motion, this conviction can be very deep, indeed. If one grows up understanding that motion is relative, then the conviction will seem as nothing more than blind superstition. When Bryce de Witt wrote a critique of Everett's interpretation, saying that "the trajectory of the memory configuration of a real observer... does *not* branch. I can testify to this from personal introspection, as can you. I simply do *not* branch," Everett responded, "I can't resist asking: Do you feel the motion of the earth?" [82, p 254]

Proponents of this objection often insist that a bifurcated/split observer in the MWI cannot in principle be considered as two separate observers, but must remain a single observer. Under this view, the Everything-happens objection becomes a problem again, since now everything really *does* happen even relative to a particular observer.

I will not respond to this objection in detail, since it is not our major concern, and, in any case, it is meaningless until one has *already* rejected the possibility of multiple superposed observers, something that has not really been argued for, but is generally simply taken as a given. To the extent that any serious argument *is* made, it is usually based on the idea of an observer being necessarily defined as a "physical system", so that even after the supposed "split", there is still only one "physical system" and hence only one "physical observer". Much is then made of the problems arising from such a presumed-singular "observer" experiencing indeterminate outcomes that he believes to be determinate (a rather extensive analysis of this approach to the MWI can be found in [15]).

However, no definition of what it means to be a "physical observer" in this sense is ever, to my knowledge, been given. It seems to be based, again, on some kind of deep conviction. The word "physical" seems to do most of the work in these arguments, but is never defined. Perhaps some sense can be made of this viewpoint, but it would certainly require a complete rejection of Everett's servomechanism assumption, since it requires the identity of a conscious observer to be defined in a manner that is clearly independent of its formal structure or its operation as a servomechanism,

and instead asks us to define the observer in terms of some ill-defined and mysterious "physical" walls that separate and unify subsystems. This seems a straightforward rejection of servomechanism equivalence, and probably wavefunction realism, as well (where are these physical "walls" in the formal structure of the wavefunction?). So, while there may be an interesting debate to be had here, it would take us far enough afield from Everett's starting point—or at least my interpretation of it—as not to be particularly relevant to our concerns here.

3.1.3 The Problem of the Preferred Basis

The problem of the preferred basis is probably the Born rule objection's major competition for most popular objection to the MWI. While our focus here is on the Born rule objection, and not the preferred basis problem *per se,* the general issue of the preferred basis is so important and pervasive in any discussion of Everett (or indeed of quantum theory more generally) that we need to have a firm gasp on it, as I do not really think it is possible to understand the probability problem without also coming to grips with the preferred basis problem.

As we have seen, outcomes can only really be defined in quantum mechanics relative to a "basis set" of states. In a sense, this basis set is the language one uses to describe the wavefunction. It is thus an analytically arbitrary choice—the wavefunction is the same wavefunction whichever language we choose to describe it in. For instance, if my basis set consists of the two states 0 and 1, I could speak of another state as being a superposition of these two states, consisting of the 1-state in some proportion, and the 0-state in some other proportion. However, this very same "superposition" might, under some *other* basis set, be itself one of the basis states, and then it would be the 0 and 1 states that would be the superpositions. When we make a measurement, however, we see a result which is one of members of a *particular* basis set, which seems to be determined by the kind of measurement we choose to make. Yet, if wavefunction realism is correct, no particular outcome resulted, since there was no collapse. Hence, the superposition of outcomes can still be understood in terms of any arbitrary basis.

The problem for MWI is that the interpretation of the universe as branching multiple worlds thus requires a *preference* for a particular kind of basis set, since the split happens according only to the preferred basis. Yet, analytically, the basis one chooses is completely arbitrary, having no more fundamental significance than one's choice of how to draw one's axes in a coordinate system.

Assume we have a photon, for instance, in the following superposition of horizontal $|0\rangle$ and vertical $|\tau/4\rangle$ polarizations (where τ is the circumference of the unit circle):

$$\left|\frac{\tau}{8}\right\rangle = \frac{1}{\sqrt{2}}|0\rangle + \frac{1}{\sqrt{2}}\left|\frac{\tau}{4}\right\rangle \tag{3.4}$$

If we believe in collapse, we could choose to say that, when we measure the polarization of the photon, the wavefunction collapses and we see one of the two superposed possibilities (horizontal or vertical). However, the situation is not as simple as this. The vertical $|\tau/4\rangle$ system (one of the individual systems in superposition) can in turn be considered a superposition of $|\tau/8\rangle$ and$|3\tau/8\rangle$ systems:

$$\left|\frac{\tau}{4}\right\rangle = \frac{1}{\sqrt{2}}\left|\frac{\tau}{8}\right\rangle - \frac{1}{\sqrt{2}}\left|\frac{3}{8}\tau\right\rangle \tag{3.5}$$

While the $|3\tau/8\rangle$ system can be considered to be:

$$\left|\frac{3}{8}\tau\right\rangle = \frac{1}{\sqrt{2}}|0\rangle - \frac{1}{\sqrt{2}}\left|\frac{\tau}{4}\right\rangle \tag{3.6}$$

So the quantum state can be analyzed into superpositions differently, depending on what we are measuring. In this case, we can choose to analyze the quantum state in terms of the orthogonal states $|\tau/4\rangle$ and $|0\rangle$, or in terms of $|\tau/8\rangle$ and $|3\tau/8\rangle$. So long as we stick to the unitary evolution of the wavefunction, it is arbitrary which point of view is taken. There is no real truth as to which states are the real "primitives", with other states being compositions of these primitives.

However, when a measurement is taken for an observable represented by an Hermitian operator that is diagonalized in one of the bases (e.g., the horizontal-vertical operator) then we will observe, as a result of the measurement operator, *one* of the basis states in this diagonalizing basis, *not* a superposition (at least not with respect to *that* basis). However, if we apply the operator for another basis (e.g., the eighth/three-eighths-τ operator), then we will see one of *its* basis states, and not a superposition (with respect to *that* basis).

As a result, it is very difficult to make any purely "statistical" interpretation of quantum mechanics work (although there are still attempts to save this idea; see [11]). The general idea of such interpretations is that the ensemble of superposed states—or "worlds" or "histories"—is not fundamentally different in character from the statistical ensembles we use in classical statistics and statistical mechanics. The quantum wavefunction is not thought to describe an actually existing single system, but an abstract ensemble of many possible systems. While this perspective works for classical statistics, the problem of the preferred basis makes it problematic when applied to quantum mechanics. Our choice of which measurement to make will affect how the wavefunction gets partitioned into members of the so-called statistical ensemble. So the superposition cannot be merely a statistical ensemble, as in classical mechanics, since our choice of measurement can affect what *are* the members of the ensemble. We could not have been in only one member of the ensemble all along, if what member we are in could have changed if we had chosen something different to measure. What this means is that the entire ensemble partakes in the dynamics of the system. In classical statistical mechanics, each member of the ensemble has its own self-contained initial conditions and time

evolution, following the mechanical laws, with no need to "refer to" the other members. In quantum mechanics, the ensemble members—if we try to think of them like that—*do* somehow seem to refer to each other to figure our how to behave (they interfere with each other). In fact, as we have seen, it is thoroughly ambiguous from the mechanics alone just what these "members" of the "ensemble" are.

From the mechanics alone, there is no problem with any of this. The wavefunction evolves in time according to its unitary law, without any need of being analyzed into subsystems, and without any need of a basis being chosen as preferred (since there is no need to view any of this evolution in terms of "measurement", or even as divided among subsystems). The reason we have a problem is that we *do* perform measurements, and our results are observed in terms of certain bases that seem to be preferred, and not others. After all, if we measure position, we see a resulting position, not a superposed smear of positions. Yet, the mathematics gives us no reason to think that the measurement process will prefer an analysis in terms of a position basis, rather than, say, a momentum basis. This does not even mean that "position" is universally a "preferred basis". We can also set up a different experiment to measure (the non-commuting observable) momentum instead, and then *that* will be the preferred basis. Yet the global state that results from either a position or momentum measurement—considering unitary evolution alone—can still be analyzed in terms of either basis. Mathematically, there is nothing about a "position measurement" that means that the position basis will be preferred. Our very labelling of it as a "position measurement" is entirely synthetic, and largely *a posteriori*, since experience has just told us that certain experimental preparations seem to yield results in terms of position. The mathematics alone does not seem to help in any way in telling which basis will be preferred, and hence what kind of "measurement" we are performing.

This is a general issue in quantum foundations, that needs to be addressed by any interpretation. However, it is probably most acute for the Everett interpretation, since the Everett interpretation seems to go further than any other in insisting that unitary evolution alone is enough, in principle, to explain our experience. Advocates of the preferred basis objection to MWI will protest, however, that Everettians *are* adding something to unitary evolution: they are adding the choice of a preferred basis, without which, there is no way of deciding *how* a measurement will "split" the universe into separate "worlds".

Response There is a consensus in the MWI community that the problem of the preferred basis has been successfully addressed [70, 183, 184], although, more generally, not everyone is convinced [85, 208]. I have some sympathy with this consensus, although I do not think it is quite the slam-

dunk that many MWI proponents feel it is. Since the preferred basis debate is not our main focus here, I will not discuss it at great length, but will give an outline of my own perspective on the issue (and these ideas will be developed in greater length later, but in the context of the Born rule objection).

The general view—amongst those who believe the preferred basis problem to be resolved—is that environment-induced decoherence explains how a basis gets "chosen", and how that basis can define how the wavefunction gets split into worlds. The idea is that "world-splitting" is not a literal splitting mechanism, but rather an artifact of decoherence. Since the off-diagonal elements in the density matrix are approximately zero, there is no effective interference involved in the observation.

One could still try, however, to make the preferred basis objection stick, as there is still no *absolute* preference that can be given to this approximately-diagonalized basis, since there is no compulsion to define measurement in terms of such a decoherence-based nearly-diagonal basis. Indeed, there is no rule given by the wavefunction itself as to how fine the coarse-graining of the Hilbert space needs to be in order to deliver the "right" set of basis vectors, since the model itself is continuous, and could, in principle, deliver a continuum of such vectors. Stapp argues this view when he states that

> ... a quantum universe evolving via the Schrödinger equation does not seem to be able to pick out a discrete set of preferred directions (or subspaces) in Hilbert space that specify quasi-classical properties that can be assigned definite "possessed values" with well defined probabilities. [208, p 14]

But doesn't environment-induced decoherence do just that, by picking out the appropriate quasi-classical basis? Not quite, since according to Stapp,

> This decoherence mechanism eliminates certain interference effects, but it does not solve the basis problem... Specifically, the effect of environmental decoherence is to reduce a typical state...to a mixture of [narrow Gaussian wave packet] states [where the density matrix for the system is traced over the degrees of freedom of the environment]... None of these states is orthogonal to any other.... There will be a continuum of these Gaussians... and they overlap: *i.e.*, they are not orthogonal... Since no one of this continuum of states is singled out from the rest by the Schrödinger dynamics, and they jointly span a large part of the Hilbert space in question, one does not immediately obtain from the Schrödinger dynamics plus decoherence the needed denumerable set... of orthogonal projection operators, or the denumerable set of orthogonal subspaces that they specify. [208, p 14]

It is tempting, if one has finitistic philosophical tendencies—as I confess to having—to dismiss this as an argument based on completed infinities, and therefore only of interest to the infinitist. Infinitistic arguments are generally dubious—at least to us finitists—since there is no reason, from any empirical data, to suggest that individual variables in nature literally posses an infinite amount of information. The tendency to be sceptical of completed infinities could be called:

Principle 3.5. *The Principle of the Discrete Default: if an objection to a theory depends solely on completed infinities or the continuum—meaning that the objection disappears when completed infinities, such as infinitesimals, are replaced with limits on discrete processes—then we should reject it (until such time as the objection has been re-constructed without invocation of completed infinities).*

Taking this approach avoids getting bogged down in the endless "paradoxes" that completed infinities have to offer. Such paradoxes may be fun for true believers in the continuum—and even sometimes for the rest of us—but if they can be replaced for all practical purposes with limits on discrete processes, then there is no necessity to be seduced by their charms. Further, if one cannot perform such a replacement, then any debate will necessarily devolve into a he-said/she-said debate between infinitists and finitists, unless it can be shown that there are actual, definable and meaningful consequences of taking the infinitist point of view (something which I believe rarely, if ever, turns out to be the case). Any such debate, I believe, must give the default assumption to the finitist, since all actual measurements and observations ever made in the real world are coarse-grained at some resolution or other. No one has ever *actually* measured a real-number-valued variable in nature, so the notion that such things exist in the real world, containing within them an infinite amount of information, has no empirical support, and is perhaps of questionable rational coherence, in the first place.

This is a moderate form of finitism which is not a revolt against the calculus, nor even against all use of literal infinities. Recall that there are no true "completed infinities" even in the integral calculus, as infinity comes into the picture only in terms of limits. And no practical use of the continuum of real numbers actually requires that each number contain a literally infinite amount of information. Any need for true real numbers can always be replaced by a discrete system, with rational numbers that, when necessary, can be made to approach real numbers, *in the limit*. It is from this moderate finitist standpoint that I am suggesting we judge arguments, like Stapp's, that are based on the continuum.

However, Stapp's continuum-based objection *can*, in fact, be replaced with a discrete limiting version, so even a moderate finitist—sceptical of the continuum, who has adopted the assumption of the discrete default—will still have to take Stapp seriously. Instead of the above, we can reword Stapp's objection in discrete language, without reference to the continuum:

> This decoherence mechanism eliminates certain interference effects, but it does not solve the basis problem. Specifically, the effect of environmental decoherence is to reduce a typical state to a mixture of an arbitrary and denumerable number of narrow, coarse-grained Gaussian wave packet states, where the density matrix for the system is traced over the coarse-grained degrees of freedom of the environment. No matter how narrow the decoherence determines these Gaussians to be, we can always choose a coarse-graining of the Hilbert space fine enough to ensure that the packets overlap: *i.e.*, that

> they are not orthogonal. Since no one of this denumerable number of states is singled out from the rest by the Schrödinger dynamics, and they jointly span a large part of the Hilbert space in question, one does not immediately obtain from the Schrödinger dynamics plus decoherence the needed denumerable set of orthogonal projection operators, or the denumerable set of orthogonal subspaces that they specify.

This is essentially Stapp's argument, replacing his invocation of the continuum with discrete language—and it seems to me that the argument comes through this transformation unscathed. I believe there may still be a problem with Stapp's argument, depending on how one interprets him, but it is not his invocation of the continuum. I said earlier, in §2.5.4, that one could always object to the idea of interaction with the environment as reducing interference to zero, since one could always try to claim that an observer might still be able to read off complete information from the environment about the complex environmental state (such as light scattering) that is correlated with her observed state. But this would be a massive amount of information. To such an observer, the "near-zero" off-diagonals would no longer be insignificant, but relatively "large" values (from the point of view of her brain with its massive information processing abilities). Such an observer would be usefully employing an extreme fine-graining of the Hilbert space of the environment, effectively creating a basis with a large number of overlapping wave packets, which could no longer be described as "virtually orthogonal".

Does this mean, then, that to solve the preferred basis problem, decoherence is not enough, but needs to be supplemented by synthetic *a priori* considerations?—for instance, that human observers have only finite information processing capabilities, that their sensory organs are capable only of observing variables of a certain dimension and structure, and so on? Well, yes, synthetic facts about human cognition and perception must be factored in; but no, this does not *exactly* mean that decoherence is inadequate to select the preferred basis. In fact, decoherence, when used to select a preferred basis through interaction with the environment, *is already* a synthetic *a priori* criterion, and will factor in exactly the relevant facts about human cognition and perception mentioned above, so long as the factoring of the Hilbert space into observer, observable and environment is done correctly—and the very use of decoherence as a tool implies that this has been done. Unfortunately, proponents of decoherence tend to gloss over this requirement, paying only lip service (if that) to the observer subsystem, and focussing mostly on the observable and environmental subsystems.

Recall from §2.5.4 that selection of the preferred basis, $\{|b_1\rangle, \ldots, |b_n\rangle\}$ by environment-induced decoherence can be viewed as the requirement that the projectors for the preferred basis, $\{\, |b_1\rangle\langle b_1|, \ldots, |b_n\rangle\langle b_n| \,\}$, commute with the interaction Hamiltonian, $\hat{H}_{int}$, that governs the dynamical interaction between the observer and her correlated complex environmental variables:

$$\forall k : [|b_k\rangle\langle b_k|, \hat{H}_{\text{int}}] = 0 \tag{3.7}$$

These two models constitute a model of how the observer subsystem interacts dynamically with the environment subsystem, for a given measurement, or measurement type. As such, it must assume some kind of model of the observer, and is at least nominally a synthetic *a priori* measure. The commutation requirement essentially tells us that the measurement process is reversible, given the interaction Hamiltonian. Of course, our model would have to be specified to an incredibly fine precision to track all the interaction effects with the environment well enough to deem the measurement process reversible. In fact, a key point about environment-induced decoherence is that the environment is hugely complex, and the information that has triggered the decoherence is lost in the chaotic complexity of the environment. If the environment were simple enough—meaning its variables were trackable by the observer—one could imagine the observer actually collecting all the relevant information required to put things back the way they were before the "irreversible" interaction took place.

Clearly, for the idea of environment-induced decoherence to make sense, the environment must be specified at a resolution, n, that is *high* enough so that the relevant observer-environment interactions are effectively irreversible. On the other hand, n must be *low* enough that it cannot be used to encode the relevant information (the mutual information between the observer and complex environmental variables) that would be required, in principle, to reverse the measurement process and permit recoherence.

Recall that if we had enough fine-grained knowledge, even of *classically* irreversible processes, like an egg's falling and breaking, then those processes would be rendered reversible. The fact that all the King's horses and all the King's men cannot put Humpty Dumpty back together again is technically a fact about the dimensionality and information capacity of the sensory organs and brains of men (and the fine motor control of horses), not a brute fact about the mechanics of breaking eggs. It is thus technically a synthetic *a priori* fact about the world, not an analytic one. However, the fact that humans lack the information gathering and processing capacity to reverse egg-breaking is blindingly obvious, so we do not question the practice of considering egg-breaking to be an irreversible physical process, even though this is not a fact recoverable from the formal egg-breaking mechanics alone (whether the egg's universe is Newtonian or quantum).

Stapp's original argument is essentially that, for any given orthogonal basis $\{|b_k\rangle\}$ that is supposedly fixed by decoherence, the dimension n of the preferred basis defines a resolution (coarse-graining) of our model of the observable that is arbitrary, and there will always be finer resolutions at which the basis elements are *not* orthogonal (and, in fact, he argues that there are a continuum of such elements). Hence, Stapp's argument essentially amounts to the claim that decoherence only fixes the preferred basis for an observable, given a preferred coarse-graining for it. Let's call this the *preferred*

resolution problem. The question of how to fix the decoherence basis then becomes the question of how to fix n.

Stapp is correct that the wavefunction itself does not deliver the degree of resolution of the environmental Hilbert space required for us to preference a virtually orthogonal quasi-classical basis. However, Stapp is perhaps making the mistake of assuming that wavefunction realism demands a purely analytic approach, due to the analyticity of the wavefunction. But this is clearly *not* the case when dealing with the preferred basis problem, since the idea here is *not* to determine an analytically preferred basis: there *is no* analytically preferred basis (as Stapp correctly shows), but neither is there a purely analytic preferred basis *problem*, since analyticity alone does not even require factoring the Hilbert space into subsystems. The preferred basis question is strictly a synthetic *a priori* question: it asks us to deliver a preferred quasi-classical basis *for* a given factoring of the Hilbert space, from the perspective of a given observer, of a specified (in our case, human) kind. Unfortunately, as physicists tend to be wary of the idea that they have to include themselves in the system—perhaps under the mistaken notion that this constitutes anthropocentrism—it is all too common for decoherence to be talked about *as if* it were strictly something that takes place between the environment and the subsystem that is being measured. But this will not work, as we have seen, if the purpose is to solve the preferred basis problem.

Hence, it is completely valid—and, in fact, necessary given the problem posed—to take as given certain facts about the sensory organs and information processing capabilities of the observers in question. To say, as Stapp seems to be, that only purely *analytic* facts about the wavefunction may be taken into account is tantamount to saying that when asked a question that assumes X, we must answer it without assuming X, and without any reference to X (which would be irrational on the face of it). Since it is an *a priori* synthetic question, selection of the preferred basis will require an *a priori synthetic* analysis[21].

Later, we will see that Wallace [227] cites the preferred resolution problem, as well, but in an effort to support the MWI against the Born rule objection, rather than attack it with the preferred basis objection. His argument is essentially that even though decoherence fixes the basis, it does not fix the coarse-graining, and therefore it does not fix the number of branches created by a measurement's splitting event. Hence, argues Wallace, the number of branches is thoroughly ill-defined, since the process of branch-splitting will ultimately be continuous (although, as in Stapp's argument, we could dispense with direct appeals to the continuum here). We will see later in this

[21] But still, for us, an *analysis*: recall from Ch. 1, that the idea of synthetic analysis is not a contradiction—I am not, by using synthetic *a priori* methods, positing the existence of unknowable essences—and indeed given psychophysical parallelism and servomechanism equivalence, we can assume that synthetic *a priori* truths, for us, are ultimately analyzable, in principle; it's just that doing so in full would require a complete analysis of human cognition, which is currently beyond our means.

chapter that this assumption is ultimately what allows Wallace's attempted Born rule proof to go through (as it allows him to assume that branch-counting can be discounted). However, if there is a preferred resolution of the basis, then decoherence should be able to fix the preferred basis, given full knowledge of all subsystem Hamiltonians, plus the interaction Hamiltonian, $\hat{H}_{\text{int}}$. As a result, neither Stapp's nor Wallace's arguments will work, since a preferred resolution will yield a discrete set of basis elements, and we will not be able to drop down to a finer (let alone continuous) grain.

We are now in a position to state our criteria for a preferred basis. This must, in principle, be defined for a *particular* measurement situation (although generalizations can be made later, to yield measurement kinds). It will assume a factoring of the Hilbert space into (at least) three subsystems: environment, observer and observable.[22] The environment subsystem is whatever is not included in the observer and observable. The observer is whatever can take on conscious states that can be sequenced according to synthetic unity. The observable is whatever becomes correlated (*i.e.*, comes to share mutual information with) the observer in the act of observation (or measurement).

The result is that the entire measurement system can be viewed as a communication of the mutual information from observable to observer, with the environment (consisting of the entire rest of the universe) acting as the communication channel. Before observation, the observer shares no (or partial) information with the observable. After observation, the observer shares full (or more) information with the observable. Because we are not viewing the channel as merely the obvious paths that the information could take, but are including even tiny chaotic effects from incidental light scattering from the environment, and so on, it may not be obvious that information has been transmitted. Because we are assuming that quantum probabilities are objective (non-epistemic) chances, mutual information is shared so long as the observer *somehow* encodes the information (in other words, it may be encoded in a way that the observer is not conscious of). However, since we are using synthetic unity to define our observer subsystem, it must be encoded in such a way as to result in a *unique* conscious state of the observer.

Crucial to the whole idea of environment-induced decoherence is the fact that the environment is not simply a passive medium through which this information flows. In the act of transmitting the message, the medium also stores the information. However, because of the complexity of the environment, the information is stored in an irreversible form (*i.e.* it would exceed the information processing capacity of the observer to re-extract this information from the environment).

Take the decoherence version of the cat experiment as an example. Because we are taking decoherence into account, we cannot assume a black box, so (unlike in the original thought experiment)

[22] Although I will continue to talk about a single $\hat{H}_{\text{int}}$, it is worthwhile noting that when we factor a system into three subsystems like this, we can decompose the interaction into three pair-wise interaction Hamiltonians [211]: $H_{observer-observable}$, $H_{observer-environment}$ and $H_{observable-environment}$.

incidental light scattering on the outside of the box *is* affected by whether the cat is alive or dead inside (it is a potential communications channel for the dead/alive information). Assume that, at least for some brief moment after the hour is up and the cat killed (or not), the observer has *not* received from (does not share with) the cat any dead/alive information. At this point, the cat is in a superposition of states, relative to that observer. Now, at the next moment, assume the light scattering off the box happens to carry dead/alive information from the cat to the observer. But what *is* the "observer"? What if we chose a factoring of the Hilbert space such that the observer was defined as just the person's *brain*? Is that adequate to ensure that the decoherence commutativity criterion will pick out a preferred basis? No, because the dead/alive information could (in principle) travel to various neurons, and affect their electrical potentials or synaptic weights, without having any effect on the observer's *consciousness* on any level. This would mean that, if we are using synthetic unity as our criterion for superposition, that the observer has still not received the relevant information, and the cat is still in a state of superposition. Thus, choosing a factoring of the Hilbert space, in terms of physical objects (even brains) can never be enough to pick out a preferred basis—specifically, it cannot tell us what kind and level of coarse-graining to use. The coarse-graining in the above example of light scattering and neurons is too fine. Such fine details of the neurons' functioning are not significant to the consciousness of the observer, and thus must either fall into the purview of the environment, or be taken care of as interaction. Hence, each possible result of an observation *must* define a unique conscious state. (Imagine instead that information from light scattering affected the neurons in such a way that there were *two* conscious states defined by the resulting basis that correspond to the *same* measurement outcome. Now the coarse-graining of the basis will be too coarse.)

These considerations force on us a *particular* level of coarse-graining for a *particular* basis, given a factoring of the Hilbert space that separates out the observer. The result is that there is no viable notion of a continuum of basis states (*a la* Stapp), nor of a continuous branching process (*a la* Wallace). Every measurement (or observation) will result in a relatively well-defined (as well-defined as the personal identity and situation of the observer allows) and discrete set of basis elements. Furthermore, there is a lower limit on the coarseness of scale of the preferred basis. The extremely fine-grained precision required for Stapp's (or Wallace's) argument to work is no longer possible. To justify such a fine-grained model of human observation, in a decoherence framework, would require positing such massive collection of information from the environment as to exceed human information processing capabilities. And even if the exact limits of human processing are debatable, it is scarcely debatable that there is *some* limit, which still means that we cannot push the resolution to an arbitrarily fine degree.

This does not, of course, tell us exactly where or how to draw the line, although there have been attempts at estimates for some of the relevant processes. It appears, for instance, that the above example of an overly-fine model of neural dynamics is *not* a realistic one (although it was not intended to be, but was posited merely for illustrative purposes). Tegmark [211] is one of the few writers on decoherence to give synthetic *a priori* considerations their due attention. He has estimated the decoherence timescales for neural firings ($\sim 10^{-13} - 10^{-20}$ seconds) and shown that they are orders of magnitude finer than the corresponding dynamical timescales for things like low-level cognitive events ($\sim 10^{-3} - 10^{-1}$ seconds).[23] Hence, by the time we reach the time-scale of conscious events, any relevant neural firings have decohered. Thus, the brain's functioning cannot rely on quantum effects and superpositions of neural states. Schlosshauer [187] has expanded on these results for networks of neurons.

To summarize my response to Stapp: the relevant factoring, and coarseness of grain, of the Hilbert space is forced on us by the nature of the observer. Hence, the Hilbert space resolution is fixed *not* merely by the requirement to produce virtually zero off-diagonals/interference, but by the information processing capacity of the observer. It is perfectly feasible that a creature could exist that does perceive our quasi-classical objects as smeared out with interference effects—perhaps there are possible creatures who would consider us to be just such creatures ourselves, from their point of view. It is futile, therefore, to expect an analytic proof for the existence of a preferred basis, without taking into account the consciousness and perception of the observer. (And we will see later that the same can be said for the probability problem.) The exact structure of the observables (including their dimensionality) is determined by the nature of this information processing and the nature of the sensory organs.

An objector might still argue that we have failed to actually state exactly *what* these synthetic constraints are. However, this is a very weak response. The point of the preferred basis objection is surely not that we cannot actually *calculate* an exact basis in practice; rather, the point is merely whether it should be possible to do so in principle, given complete (or at least sufficient) knowledge of the system state. Since the observer is part of the system, there is no reason not to suppose that the decoherence basis could not in principle be calculated. Since for *any* observer, there *will* be such cognitive and perceptual limits and constraints, the mere fact that we do not currently understand

[23] In addition to regular neural firings, Tegmark shows this for kink-like polarization excitations in microtubules. This is motivated by speculations by Penrose [154, 153, 152] that microtubules in the brain can process information while maintaining quantum coherence, and hence acting as a kind of quantum computer, while exploiting gravitational effects (Penrose hypothesizes that the brain's processing is non-algorithmic, so he cannot simply postulate that brains are quantum computers, since quantum computers—while differing from deterministic computers in the complexity of their algorithms—are still thoroughly algorithmic in nature. Penrose's speculations are just that, as there is no real evidence that microtubules play this role, and Tegmark shows that it is highly unlikely they would be physically capable of it.)

human cognition and perception—in the incredible depth and detail that would be required to precisely derive the relevant decoherence basis—in no way implies that there is a "preferred basis problem" for the MWI, given that our nature as observers, in combination with the wavefunction, must of necessity demand a preferred basis of some sort.

Returning to the cat experiment, in cases where tiny effects in the environment can affect (or encode information about) whether the cat is dead or alive, then the correlated environmental states will be nearly orthogonal to each other, and the system state will be *nearly* decoherent. There is, however, nothing about the wavefunction *analytically* that determines this—we need to factor in that our visual system cannot perceive a higher-dimensional live-dead cat, and that we cannot resolve the environmental effects to such a fine resolution as to trace the effects of light scattering on the cat, which would require a massive amount of information about the environment.

Stapp might argue that we have not explained why some *other* basis, not virtually-diagonalized, should not define our "measurement", for there is no reason why the quasi-classical virtual disappearance of macroscopic interference should be a *required* feature of measurements.

Imagine, then, a creature in the universe (or multiverse, if you prefer) that *can* observe superposed live-dead cats. Such a creature would have to have higher-dimensional sensory apparatus than ours, and its brain would likely have to have far greater information capacity than ours, but there is no reason to assume that such a creature does not exist out there, in some corner of the uni/multiverse. Perhaps this creature would *not* observe either a dead *or* a live cat when looking at the opened box, but rather would directly observe a superposition of dead/alive, as directly as we observe a rainbow. Before the box is opened, this creature would receive the same environmental which-path bits of information that we do. For us, such information reduces our off-diagonals to zero and removes interference. But for the higher-capacity creatures, their off-diagonals are *not* reduced by this same information, coherence is maintained, and they can directly observe the superposition. They will also directly observe that the human observer is in superposition. They would not consider such states to be "superpositions", however, or to be any more mysterious than we consider rainbows to be.

On the other hand, these creatures will still have sensory organs and informational capacities that place limits and constraints on *their* observables. So they will have their own degree of resolution that will impose its own constraints on which observable will display interference and which will not. So there will still be superpositions from the point of view of such a creature—which it will be able to detect, by doing statistical experiments just as we do—but these states will be at a higher dimensionality and level of information processing than human observables. The possible existence of such creatures disproves the contention that the mathematical analysis of environmentally-induced deco-

herence can, applied to the wavefunction, select out a preferred basis. Synthetic *a priori* constraints on the observer must be taken into account.

These same constraints that determine the effective basis also tell us how the wavefunction will "branch" into perspectival worlds. The worlds are the terms in the superposition, according to the preferred basis, such that each world contains a separate conscious observer. If a single branch contains two incompatible observers, then our coarse-graining is not fine enough. If it splits a single observer across worlds, then the grain is too coarse. We could pretend to split things into entirely different kinds of "worlds", not tied to the perceptual characteristics of the observer, if we chose to—since this would be allowed analytically—but since the splitting of the system into so-called worlds is only a matter of perspective in the first place, we really are forced to choose our basis on *a priori* synthetic grounds, dividing the system into such worlds where separate consciousnesses inhabit separate worlds. This is, in fact, part and parcel of what (in an Everettian context) we must surely mean by "world".

3.2 The Born rule Objection

Current consensus is that the Born rule objection is by far the most serious of the objections to the MWI. Even those defenders of the MWI do not believe it necessary to actually prove the Born rule from MWI assumptions, will have to admit that such a proof would greatly strengthen the credibility of the interpretation.

The Born rule tells us that the probability of a certain result in the wave function is equal to the square of the norm of the amplitude of the wavefunction at that point. Since there is currently no purely analytic proof of this rule from the quantum wavefunction itself, it remains an independent postulate of quantum mechanics. However, Everett's stage 1 proof shows that, so long as we assume amplitude dependence—that the amplitude for a state is all that matters to its probability—then a norm-squared measure follows. In fact, this measure is mandated so long as we assume the Gleason noncontextuality postulate: that the probability of a state is dependent only on that state, and not on what basis it is considered to be a part of. In fact, Gleason's proof can be formulated by first proving dependence on amplitude [135] from noncontextuality, so it seems that noncontextuality is closely related to amplitude dependence (the former implying the latter).

So it is not the exact form of the Born rule that is at issue. What is at issue is the more general idea that it is amplitudes that matter. Or, to take the Gleason perspective, that basis membership does *not* matter. The Born rule objectors point out that Everettians claim many-worlds to be a natural consequence of taking the quantum formalism seriously on its own terms. If that is true, they say, then the MWI should be able to deliver the Born rule without *any* assumptions. However,

I have already argued that it is not reasonable to insist that the probabilities associated with an observation be calculated purely analytically without any model of what observation is (and no one is reasonably asking for a complete analytic model of observation).

However, the Born rule objection at its most cogent is not really about demanding a formal Born rule proof. Proponents almost always view branch-counting as a necessary *a priori* counting method, required for the MWI to work. Hence, Gleason's theorem is actually a *reductio ad absurdum* of the MWI, from this perspective. The MWI, they argue, demands branch-counting, since Everett contends that all the branches or worlds are equally real. In addition, he directly supports the assumption of branch counting in his stage 2 proof.

And indeed, argues the Born rule objector, how can the branches be equally real, if they are not equally probable? Hence, the MWI demands a world-counting measure. Yet the empirical evidence tells us that the true measure is an amplitude-counting measure. And while it is possible that an argument could be made for Gleason noncontextuality from wavefunction realism, this would still conflict with the many worlds interpretation, since noncontextuality demands an amplitude-counting measure and the MWI demands branch-counting. The MWI is therefore in direct conflict with both empirical evidence *and* Gleason noncontextuality. Therefore, Everettian branches simply cannot be "equally real", as Everett claims. The whole notion of equal reality of branches conflicts with the structure of Hilbert space, and thus the many worlds interpretation cannot be the whole story.

3.3 Responses to the Born rule Objection

3.3.1 The Double-standard Response

Some supporters of Everett feel that many of the most common objections, including the Born rule objection, are a double standard [213]. Other interpretations are not required to derive the Born rule, so why should the MWI? The justification for the double standard seems to come down to this: other interpretations do not make such sweeping claims as "the wavefunction and only the wavefunction", and therefore have less need to derive the Born rule, since they do not claim that the wavefunction stands on its own in the first place. Many of them, in fact, actually suggest subtle, or not so subtle, changes to the very dynamics of the wavefunction (see [89][19])—the Copenhagen Interpretation does this as well, in a sense, relying as it does on the presupposition of classical objects lying completely outside the wavefunction.

The MWI must answer to a higher standard than these other interpretations, it is said, because it is making pretensions to do more, and its claim to such metaphysical elegance is supposedly the main attraction for accepting its highly counter-intuitive results in the first place.

What this view ignores, however, is that the Born rule is not itself a modification to the dynamics of the wavefunction. It is rather a rule for computing probabilities *from* the dynamics. It is possible that a deeper understanding of the wavefunction—or even just a deeper understanding of probability theory—might yet be forthcoming, one that would in no way conflict with the dynamics of the current theory, but which would justify why the Born rule is the appropriate way to compute probabilities, given the dynamics. While such a tidy justification is not currently to be had, it is not clear that this deeper understanding should be required in choosing the best interpretation of current theory. We will see in later chapters that this is especially true, given that the interpretation of probability is itself controversial. If the *a priori* derivation of a probability rule from the wavefunction depends on the choice of the best interpretation of probability theory, and there is no general consensus on the interpretation of probability theory, then should the MWI be held accountable to the inadequacies of probability theory? Surely, then, the correct response to the objectors is simply to suggest everybody wait until probability theory is properly figured out. This is why, in fact, I will end up spending a great deal of this dissertation examining the foundations of probability theory, at least the theory of objective probability or chance.

Also, the extent to which one feels the MWI should be held to a higher standard depends heavily on how one views the MWI in relation to Occam's Razor. For those who feel that positing multiple universes is an extremely drastic move, the fact that the Born rule does not arise out of the dynamics alone is a great problem. For others who may see little problem in positing multiple universes if the resulting system is in any way conceptually simpler, there is considerably less need to deliver the Born rule before taking the MWI seriously, especially if the possibility of a consistent justification arising out of the dynamics has not actually been ruled out.

Ultimately, whether or not one believes that the Born rule objection is fatal for the MWI or not, clearly we can all agree that if the MWI could deliver the Born rule, it would be a much more successful interpretation, so the probability objection clearly cannot be outright rejected. Even those MWI supporters who feel the Born rule objection is a double standard will presumably support attempts to derive the Born rule within an MWI context.

3.3.2 The "What-else?" Response

This general response, championed by Saunders [185], is called the "bemusement" response by Wallace [226]. In a way, it is a variation on the double-standard response. It is a stronger position, however. The double-standard position acknowledges the problem with not having a Born rule derivation, but sees this as not the exclusive problem of the Everett interpretation. The bemusement responses are more likely to suggest that the lack of a Born rule proof is not so much an urgent

gap to be filled, but rather the expected state of affairs, and that the Born rule is nonetheless the most natural and expected measure—or at least not unexpected—and that the lack of an absolute proof is perfectly acceptable.

Going along with this response is usually some degree of recognition of synthetic *a priori* considerations: that it is a benign and necessary part of the application of probability theory to consider the nature of the observer, rather than to get stuck on finding purely analytic formulae. Saunders puts it this way:

> "... the question is no longer: What is the space of possibilities?—but: What is the space of possibilities in which we are located? That we are ourselves constituted of processes of a very special sort is in part explanatory; because our nature is in part contingent, a matter of evolutionary circumstance, the demand for explanation is blunted." [185, p.374]

I would interpret that the "blunting" of the need for explanation is essentially the fact that we may be forgiven for not having a complete analytic model of observation.

Another component of "What else?" is the fact that the Born rule is essentially nothing more than the counting of amplitudes, and amplitudes are what the wavefunction consists of. It is clear (analytically) what an amplitude is. It is completely unclear what a "branch" is. Therefore, what else would we count, other than amplitudes?

Everett points out that this whole situation is no different from the situation in classical statistical mechanics, which also lacks an actual proof of its probability measure on the phase space. The actual measure used, and confirmed by experiment, is the Lebesgue measure, which is a simple volume measure. While this seems to many the obvious measure, there is no analytic proof for it. It is a synthetic presupposition. Yet, there is no widespread concern that some kind of serious "metaphysical baggage" is added to classical mechanics here, since the mechanics itself is not being modified in any way. No extra constituents of reality are being posited here.[24]

If there is something else, other than amplitudes, that can reasonably be considered as inherent to the wavefunction, that we can count or measure in some way that yields a reasonable probability rule, then the "What-if?" respondents ask us to present it. Otherwise, there is nothing pathological about accepting amplitude-counting. Predictions in any theory need some sort of interface between the empirical acts of observation and the analytic model. The model will never deliver this interface entirely on its own (although, as always in science, we are guided by William of Occam). Where else in science, they ask, are empirical probabilities expected to arise entirely out of pure analysis?

[24] There are still those who puzzle over the Lebesgue measure, however, and consider it a problem in need of a solution [5, 66], and I am in no way dismissing such concern. However, if even *classical* statistical mechanics cannot justify its analogue to the Born measure, it at least calls into question the seriousness of the same lack existing in quantum mechanics.

The Born rule objectors, however, will normally bring forth branches or outcomes, or something analogous, as the alternative countable. Thus, they feel they have a good answer to the "What-else?" question. However, I argue that a true Everettian, *if* a believer in objective probability and in Everett's wavefunction realism, need not be concerned with branches, since amplitudes are the coinage of the wavefunction, not outcomes or worlds—and Everett's fundamental starting point is "wavefunction realism", not "world-realism".

Everett himself, however, does *not* hold to such a purist version of his own principle. His stage 2 proof is entirely devoted to a frequentist proof that branch-counting amounts, in the limit, to the Born rule. Thus, it seems that Everett himself, gave significant credibility to the idea of counting branches. This might not be a problem for him if he supported subjective probabilities, but most Everettians seem to hold to an objectivist notion of probability—and certainly I will be assuming objectivism throughout this dissertation. Everett himself was not explicit about his general philosophy of probability, although his tone is objectivist and he employs a frequentist approach without questioning it.

The whole "What-else"? argument generally lacks clout for subjectivists (although I will not dwell on this issue, given that I will generally be adopting the objectivist stance, anyway). For the subjectivist, the fact that amplitudes are primitive elements of the wavefunction, rather than emergent properties, is not especially relevant. If probabilities are subjective, we would in fact expect them to be (possibly) based on emergent properties. Hence, my arguments above in favour of "What-else?" are not conclusive, and only establish that *if* ontic countables are the stuff of probabilities, *and* worlds are purely synthetic/phenomenal entities, then the countables *might* be amplitudes, while they *cannot* be worlds. This is not an especially strong result, depending on your presumptions.

To further clarify and generalize this point, I will suggest that it is inconsistent to support all three of:

1. objective quantum probability,
2. perspectival branching, and
3. *a priori* branch-counting.

One can maintain *any two* of these three assumptions, but one (at least) must be done away with. If quantum probabilities are objective, and branching is purely perspectival, then the correct *a priori* rule cannot be branch-counting, for that would count purely perspectival entities and our probabilities would not be objective. On the other hand, if we accept objective probability, while maintaining that branch counting holds *a priori*, then branches must not be merely perspectival,

since we are counting them to obtain an objective probability. And finally, if we maintain that branches are perspectival, and that they are the countables, then we cannot maintain objectivity of the resulting measure, which is a count of purely perspectival entities.

It has been very common for Everettians to assume that wavefunction realism demands both objective probabilities and perspectival branching. Yet, it has also been common to accept the idea of *a priori* branch-counting, resulting in the numerous frequentist attempts to equate the Born rule with branch-counting in the limit. I would suggest that there is an inherent inconsistency in this position, one that started with Everett himself, who seemed to argue for a thoroughgoing wavefunction realism—viewing branching as perspectival—while still feeling the need for a frequentist reduction of quantum probability to branch-counting, a move that was not only unnecessary, given his wavefunction realism, but possibly even at odds with it, seeing as it implies some kind of objective reality for branches.

Whatever Everett's original intent, most of the serious defences of Everett's interpretation these days hold that worlds are a phenomenal emergent property; they are not ontic entities or even inherent elements of the wavefunction. Furthermore, Everett never explicitly advocates the idea that worlds are real entities.[25] He may have simply thought that a limit-based equivalence between the two measures would undoubtedly be helpful in satisfying his critics, or he may have felt that the countables in a probability measure should, in fact, be phenomenal and not ontic. In any case, what Everett *does* make clear, explicitly and repeatedly, is that his whole interpretation is primarily based on the assumption that it is the wavefunction and the wavefunction alone that provides the underlying ontology. Thus, I think it is reasonable to conclude that the appropriate Everettian stance is that wavefunction amplitudes are ontic, while worlds and measurement outcomes are phenomenal.

Assuming we take this stance on Everett, then I feel the "What-else?" response is quite cogent, and would agree that an actual derivation of the Born rule from the dynamics alone is not necessarily called for. There is still, however, no reason to ignore the call for a Born rule derivation. It would clearly add to our understanding of the MWI immensely, whether or not we think it mandatory. In addition, the double-standard and "What-else?" arguments clearly will not convince everyone, especially when they appeal to ideas about how our perceptions interface with the environment, as any such considerations will be met with widely different reactions in the scientific community. There are many who are automatically suspicious of any such talk. In addition, there are others who really

[25]Some may argue that he *did* imply it, since he said that the branches are "equally real", which amounts to the same thing. However, the claim that A and B are "equally real" does not in general imply that they are both "real". For instance: "Sherlock Holmes and Professor Moriarty are equally real." Or: "The beauty of Euclid's Proposition 47 is as real as the beauty of the red, red rose." Both horizons ahead and behind me are "equally real", but neither is a "real thing" in the world. They are *effects*—artifacts of my perception of real things—but not things-in-themselves. Yet I still say that one horizon is "just as real" as the other, without believing them to be fictitious. After all, they may be artifacts *qua* horizons, but that does not make them unreal or figments of my imagination.

do unapologetically support a double-standard for Everett. They find the use of multiple worlds to be so egregious—even outlandish—that they simply feel that the level of proof required must be set extremely high. Comparisons with how we handle comparable issues in statistical mechanics, and other classical contexts, will not impress them much, even if they agree that there is a valid comparison to be made. Their position is that "extraordinary claims require extraordinary evidence", and there is a feeling that Everett's claim is extraordinary enough to demand nothing short of an actual Born rule derivation.

Regardless of where we stand on the need for such a derivation, there is little doubt that Everett's interpretation would be much stronger if it had one. So we will now move on from considering whether such a proof is required, and look at the efforts that have actually been made to provide one.

3.3.3 Proof-based responses

Perhaps the type of response with the most history is the attempt to derive, or at least partially justify, the Born rule within an MWI context, using relative frequencies of outcomes. Everett's original probability argument [79, pp 460-461] as well as others [95, 83, 103, 182, 101] fall within this general category. Most such attempts, including Everett's, essentially seek to prove that the number of worlds that do *not* conform to the Born rule goes to zero in the limit of infinite worlds. However, none of the attempted frequentist proofs of this type have yet met with widespread acceptance. They are generally viewed as circular in nature, implicitly assuming amplitude dependence in the process of trying to prove an amplitude-counting rule.

Other approaches try to solve these problems by sticking with some kind of branch or outcome or observer counting, but essentially reformulating the idea of what is counted, or refining the idea of counting (or even tweaking the dynamics) so that the Born rule can be recovered after factoring in other considerations, such as decoherence effects [102, 229], scaling considerations [229, 210], or the effects of noise [139]. The entities being counted might be worlds [102], outcomes [229, 210], or observers [139]. However, in these approaches, the process of "outcome" counting has usually been significantly re-interpreted from its conventional sense, so it may be mere semantics whether we should count these approaches as branch-counting or not. My own approach has much in common with these efforts, and can be classified in this category, since although I will be rejecting the idea of counting observers, I will be doing something closely related, in counting observer *algorithms*. I do not, however, find it useful to define my kind of counting as a refined kind of branch-counting, as I find it more useful to reject the branch-counting mentality altogether.

Then, there are the responses that attempt to derive the Born rule, but do not employ any

kind of count-based interpretation of probability at all, most notably the Deutsch-Wallace proof [71, 226], based on decision theory. This is one of the few MWI-based Born rule proofs that takes a broadly Bayesian approach, although it must be noted that it does not take an extreme Bayesian (subjectivist) stance, as some of its axioms [226] are explicitly objectivist in nature (for instance, the probability of a state is effectively assumed to be determined solely by that physical state, an assumption tantamount to objectivism, as well as to Gleason noncontextuality).

Finally, there is Gleason's proof [91] itself, which is analytically the strongest of the Born rule proofs—and certainly the most influential outside of the MWI community. Solely from noncontextuality, Gleason proves that there is no possible measure other than the Born measure, without invoking infinite limits or sneaking amplitude dependence in through the back door. And the sole synthetic axiom simply assumes that the probability of a state is independent of which basis we consider it to be a member of, an assumption that is very difficult for any objectivist to deny, and (as we will see) arises naturally in an algorithmic context. I include Gleason's proof here because of its superiority (in my opinion) as an MWI Born rule proof, *not* due to its popularity, as it is generally believed to be completely irrelevant.

3.3.4 The Everett Proof

Recall that Everett's relative state interpretation is based on the fundamental assumption that the unitary evolution of the wavefunction is all there is. Start with a state, which can be analyzed as a sum of n orthogonal basis states in an n-dimensional Hilbert space:

$$|\psi\rangle = \sum_{i=1}^{n} |\psi_i\rangle = \sum_{i=1}^{n} a_i |i\rangle \tag{3.8}$$

where the elements of $\{|i\rangle : i = 1 \cdots n\}$ are the orthonormal basis states. We say that $|\psi\rangle$ is in a superposition of these basis states, where each member $|i\rangle$ of the basis is an "element" of the superposition, with an amplitude a_i (but, of course, $|\psi\rangle$ may be decomposable into other superpositions, using other bases). Measurement gives only a probabilistic result, which experimentally we know to be based on the Born rule (probability as norm-squared amplitude):

$$p(\psi_i) = |\langle\psi_i|\ \psi\rangle|^2 \tag{3.9}$$

Whereas the collapse postulate states that only one element of the superposition remains after measurement, Everett's "no-collapse postulate" states that there is no such collapse, and only the wavefunction evolution is needed. The Born-rule objectors, as we have already discussed, feel that if Everett is to make such lofty claims (that the wavefunction is all there is) he ought to be able, then, to derive the Born rule *from the wavefunction alone*.

Everett does indeed attempt to prove this. I will outline his proof here.

Everett's "proof" is really two proofs, each of which has a very different flavour and purpose. I will call them stage 1 and stage 2.[26]

Theorem 3.6. ***Everett's Probability Theorem, Stage 1:*** *the only $p(A|\psi)$, that are functions of the normalized state $|A\rangle$ given wavefunction $|\psi\rangle$, that are preserved under additivity, are the norm-squared amplitude functions:*

$$p(A|\psi) \propto |\langle A|\psi\rangle|^2 \tag{3.10}$$

Proof. Stage 1 is based on a fundamental assumption:

Assumption 3.7. ***Amplitude dependence:*** the probability $p(A)$ for a state $|A\rangle$ given wavefunction $|\psi\rangle$ is solely a function of the amplitude for state $|A\rangle$ in $|\psi\rangle$.

Now, since the $|i\rangle$ basis states are of unit norm, we can just use the amplitudes (the a_i's) to calculate probabilities. Because amplitudes are only determined up to an arbitrary phase factor, anyway, we can in fact just use norm of the a_i's. The probability $p()$ of an outcome, then, will be a function of the a_i norms alone:

$$p(a_i) = p(|a_i|) \tag{3.11}$$

Now consider an m-dimensional subspace, as a subset of our n-dimensional basis (which we assume, without loss of generality, to be the first m elements of the n-dimensional basis):

$$\{|i\rangle : i = 1 \cdots m; m < n\} \tag{3.12}$$

Consider also the corresponding "subset" of our superposition, $|\psi\rangle$:

$$a\,|\varphi\rangle = \sum_{i=1}^{m} a_i\,|i\rangle \tag{3.13}$$

which is just the (unnormalized) projection of $|\psi\rangle$ onto our m-dimensional subspace. Without any loss of generality, we will assume[27] that $m = 2$:

$$a\,|\varphi\rangle = a_1\,|1\rangle + a_2\,|2\rangle \tag{3.14}$$

[26] Note that Everett himself did not divide the proof explicitly into these two stages. I am doing so because I feel the two stages accomplish very different things, and Stage I can stand on its own, so far as it goes, whether we assent to Stage 2 or not.

[27] Everett does not make this assumption, but it does not affect the substance of the proof, and is introduced here to keep things simple.

Everett now employs another fundamental assumption (which I have here rolled into the theorem itself), which is additivity: *probabilities are preserved across additions of basis states. I.e.*,

$$p(\varphi) = \quad p(1) + \quad p(2) \; = \tag{3.15}$$

$$p(|a|) = p(|a_1|) + p(|a_2|) \tag{3.16}$$

This is, of course, a main requirement for a function to be a probability measure (in addition to summing to unity over the entire space, and yielding non-negative results). Therefore, it seems that Everett is well-entitled to this assumption (and it is, indeed, usually the previous assumption, not this one, that Everett's critics are prone to attack).

Since the wavefunction dynamics are unitary, inner products are preserved (by definition of unitary evolution), and therefore so are norm-squareds. By definition of the norm:

$$|a| = \sqrt{|a_1|^2 + |a_2|^2} \tag{3.17}$$

This is simply Euclidean distance, a generalization of the Pythagorean theorem [78, Bk.1, Pr.47-8]:

$$|a|^2 = |a_1|^2 + |a_2|^2 \tag{3.18}$$

So a norm-squared measure observes additivity of probabilities, which is required for probabilities to make sense as probabilities.[28]

Everett now needs to show the converse: that there is no other function of amplitude that obeys additivity other than norm-squared. Since we can replace complex amplitudes with their norms, for purposes of calculating $p()$, it follows from normality that:

$$\begin{aligned} p(a) &= p\left(|a|\right) \\ &= p\left(\sqrt{|a_1|^2 + |a_2|^2}\right) \end{aligned} \tag{3.19}$$

and by additivity

$$\begin{aligned} p(a) &= p\left(|a_1|\right) + p\left(|a_2|\right) \\ &= p\left(\sqrt{|a_1|^2}\right) + p\left(\sqrt{|a_1|^2}\right) \end{aligned} \tag{3.20}$$

So normality and additivity together yield

$$p\left(\sqrt{|a_1|^2 + |a_2|^2}\right) = p\left(\sqrt{|a_1|^2}\right) + p\left(\sqrt{|a_1|^2}\right) \tag{3.21}$$

[28] In addition to additivity, of course, our measure must be normalized so that it sums to one over the probability distribution.

Since every argument to $p()$ above is a square root, we can re-phrase this using a "root-measure" function $p^{\surd}()$:

$$p^{\surd}(x) = p\left(\sqrt{x}\right) \tag{3.22}$$

as

$$p^{\surd}\left(|a_1|^2 + |a_2|^2\right) = p^{\surd}\left(|a_1|^2\right) + p^{\surd}\left(|a_2|^2\right) \tag{3.23}$$

In other words, the root measure $p^{\surd}()$ has the same result applied to a sum as the sum of its applications to the individual terms of the sum—*i.e.* it is a *linear* measure. It follows therefore that

$$\begin{aligned} p^{\surd}(x) &\propto x \\ &= Kx \end{aligned} \tag{3.24}$$

where K is a constant.

It follows that applying our original measure $p()$ to something is effectively the same as squaring with a multiplicative constant:

$$\begin{aligned} p^{\surd}\left(x^2\right) &= p\left(\sqrt{x^2}\right) \\ p(x) &= Kx^2 \end{aligned} \tag{3.25}$$

And we have therefore deduced that the probability measure is a function of norm-squared amplitude:

$$\begin{aligned} p(i) = p(a_i) &= p(|a_i|) = K\,|a_i|^2 = Ka_i^* a_i \\ p(i) &\propto a_i^* a_i \end{aligned} \tag{3.26}$$

□

Thus, norm-squared amplitude not only obeys additivity, it is the only measure that does so, and it is a unique measure if we assume normalization, so that the measure sums to unity (which dispenses with the arbitrary constant K). The proof above is expressed for only two outcomes, for clarity. Everett's original proof generalizes this straightforwardly to any number of outcomes.

This proof is convincing as far as it goes, but it assumes amplitude dependence: that the (unnormalized) measure of any single term in a superposition is a function of *just* the amplitude of *that* term. The resulting rule is thus a form of amplitude-counting: it assumes that only the amplitude of an outcome matters to its probability. Stage 1, then, merely proves that the norm-squared measure is the right way to count amplitudes, but it tells us nothing about why it is we should count amplitudes in the first place. It thus does not take into account the possibility that there could be some (possibly very complicated) function of the entire superposition that yields probabilities that also obey additivity and summation to unity. In fact, branch-counting is arguably just such an alternative measure.

Theorem 3.8. ***Everett's Probability Theorem, Stage 2:*** *sequences of observations of length N obey the statistics of the measure defined in Stage 1 (the Born rule) in the limit, as N approaches infinity. I.e., the set of sequences of infinite length that do not obey the Born rule are of measure zero.*

Proof. Stage 2 is concerned with the tension (and apparent contradiction) between two alternative statistics for calculating probabilities in the wavefunction: observer-counting (or branch-counting) versus amplitude-counting. As mentioned earlier, Everett implicitly condones the idea that world branches are ontic entities in this stage, which is motivated by the perceived problem that, regardless of whether we use the Born rule probability measure derived in stage I, *most observers will nonetheless not observe (and hence most worlds will not obey) Born rule statistics over a sequence of many repeated observations.*

To see why this is so, consider sequences of observations (which we can consider either to be world histories, or, like Everett, observer memory trajectories). Assume the observations are of some binary result, either $|0\rangle$ or $|1\rangle$. Given the no-collapse postulate, the result of a single observation is a superposition $|\psi_1\rangle$ of $2^1 = 2$ states:

$$|\psi_1\rangle = a_0 |0\rangle + a_1 |1\rangle \tag{3.27}$$

where a_k is the amplitude for result $|k\rangle$.

If we repeat the observation n times, each time on an identically prepared system, we end up with a superposition of 2^n states, which could be considered worlds or world histories:

$$|\psi_n\rangle = \sum_{r_1 \cdots, r_n} a_{r_1 \cdots, r_n} |r_1 r_2 r_3 \cdots r_n\rangle \tag{3.28}$$

Applying the Born measure p from stage 1, the probability $p_n^B(r_1 \cdots r_n)$ of a particular length-n sequence $\{r_1, \ldots, r_n\}$ will be equal to the product of the probability measures for each individual outcome:

$$p_n^B(r_1 \cdots r_n) = p(r_1) \cdots p(r_n) \propto |a_{r_1}|^2 \cdots |a_{r_n}|^2 \tag{3.29}$$

In general, the $p_n^B(r_1 \cdots r_n)$ values will not all be equal. Some may, indeed, be very tiny. We can also use the same measure for the probability of a set Q of m sequences $\{q_1, q_2, \ldots, q_m\}$ of length n:

$$p_n^B(Q) = \sum_{k=1}^{m} p_n^B(q_k) \tag{3.30}$$

Our probability distribution over all 2^n sequences obviously obeys the Born rule (since the Born rule generated it). However, the individual sequences themselves do not *each* obey Born rule statistics. In fact, most do not. It cannot even be said that individual Born-rule obeying sequences are more probable than individual non-Born sequences.

This is not anything mysterious; it's just probability. Consider, for example, the case where the Born rule tells us that

$$\begin{aligned} p^B(0) &= \frac{1}{4} \\ p^B(1) &= \frac{3}{4} \end{aligned} \tag{3.31}$$

The sequence {1110}, which exhibits exact Born behaviour for $n = 4$, is more common than {1010} or {0000} , but less common than {1111} (none of which are matches for Born behaviour):

$$p_4^B(1110) = \frac{3}{4}\frac{3}{4}\frac{3}{4}\frac{1}{4} = \frac{27}{256} \approx 10.5\% \text{ (obeys Born rule)}$$

$$p_4^B(1010) = \frac{3}{4}\frac{1}{4}\frac{3}{4}\frac{1}{4} = \frac{9}{256} \approx 3.5\% \text{ (disobeys Born rule)}$$

$$p_4^B(0000) = \frac{1}{4}\frac{1}{4}\frac{1}{4}\frac{1}{4} = \frac{1}{256} \approx 0.4\% \text{ (disobeys Born rule)}$$

$$p_4^B(1111) = \frac{3}{4}\frac{3}{4}\frac{3}{4}\frac{3}{4} = \frac{81}{256} \approx 32\% \text{ (disobeys Born rule)} \tag{3.32}$$

Note here that when I say "disobeys Born rule", I just mean that the sequence does not exhibit Born rule statistics; clearly one would not really draw any conclusions about the source of the sequence, and what rule it was following, based on only four data points.

Here is a point that can cause confusion: the most Born-probable of these four worlds is a non-Born world, which is fully three times more likely than the Born world. So p^B, while it is the Born probability of a sequence, tells us nothing of whether that sequence follows the Born rule or not, since the most likely individual sequence will always just be a string of whatever the most likely result is. Imagine you have an unfair coin weighted to fall heads three times out of four. If you flip it a million times, the most likely *individual* sequence is a million heads in a row, with a probability of about $1.8 \times 10^{-124937}\%$. That sounds low, but it *is* the single most probable sequence. If you had to bet on an outcome, that would be the most rational choice. The likelihood of any particular sequence that actually obeys the correct statistics, with 3/4 heads and 1/4 tails, is only about $8.9 \times 10^{-244218}\%$, which is about 2×10^{119280} times *less* likely than getting a million heads in a row. Again, this is not mysterious, just probability. But if we are not clear on this, it can cause confusion, especially when moving into a domain that is more abstract and difficult to imagine than coin flipping.

To actually determine whether we are looking at a Born sequence or not, we compare the average value to the expected Born average, using the law of large numbers.

For our example, the mean of the values in the sequence should be close to $p(1)$ in a Born world:

$$\frac{\sum_{i=1}^{n} r_i}{n} \approx p(1) \tag{3.33}$$

Definition 3.9. Define the "Born error" $E_n^B(r_1 \cdots r_n)$ as the percent error of the mean compared to the expected value for a sequence $\{r_1 r_2 \cdots r_n\}$:

$$E_n^B(r_1 \cdots r_n) = \frac{\left| \frac{\sum_{i=1}^{n} r_i}{n} - p(1) \right|}{p(1)} \tag{3.34}$$

And define $B_n(r_1 \cdots r_n)$ as the "Born measure" of a world or sequence of measurement outcomes—*i.e.,* how close a sequence comes to being a Born sequence:

$$B_n(r_1 \cdots r_n) = 1 - E_n^B(r_1 \cdots r_n) \tag{3.35}$$

The Born error should be close to 0% for a Born sequence, meaning it has a Born measure close to 1:

$$\begin{aligned} E_n^B(r_1 \cdots r_n) &\approx 0\% \\ B_n(r_1 \cdots r_n) &\approx 100\% \end{aligned} \tag{3.36}$$

Calculating B_n for ever-increasing values of n is something like what we might actually do to determine whether we are living in a Born world. And, of course, empirically we actually do find that B_n does get very close to 100% for very large n. To prove the Born rule in the limiting case, we want to show that

$$\lim_{n \to \infty} B_n(r_1 \cdots r_n) = 1 \tag{3.37}$$

The point of the current proof is to show that the Born rule follows *a priori* from the wavefunction, interpreted as a tree of bifurcating observers. Everett felt that this demanded more than simply showing that the Born rule follows from amplitude-counting plus additivity (stage 1). He felt, in addition, that whatever probability rule we end up with, for *a priori* reasons, has to ultimately square with the statistics of observer-counting. This is a fundamental assumption of stage 2 of Everett's proof:

Assumption 3.10. *The assumption of observer-counting. If the wavefunction yields b outcomes to measurement M, relative to observer O—who therefore bifurcates into b post-measurement observers $\{o_1, \ldots, o_b\}$, collectively experiencing b measurement results $\{r_1, \ldots, r_b\}$—then the probability $p(r_k)$ that observer O ought rationally to assign to outcome r_k is the outcome-counting probability $p^w(r_k)$:*

$$p(r_k) = p^w(r_k) = \frac{1}{b} \tag{3.38}$$

For our binary-outcome example, $b = 2$, this becomes

$$p(r_k) = p^w(r_k) = \frac{1}{2} \tag{3.39}$$

I have stated this explicitly as an arbitrary assumption, to highlight its role in the proof, but Everett thought it was a necessary part of trying to assign probabilities to sequences in superposition:

> In order to establish quantitative results, we must put some sort of measure (weighting) on the elements [sequences/worlds] of a final superposition [of sequences/worlds]. This is necessary to be able *to make assertions which hold for almost all of the observer states* described by elements of a superposition [i.e, worlds]. We wish to make quantitative statements about the *relative frequencies...for a typical observer state*; but to accomplish this we must have a method for selecting a typical element [or world] from a superposition of orthogonal states. [79, p460, emphasis mine]

Everett assumes that a probability measure on worlds (or observers or outcomes) must allow us to make "assertions which hold for almost all of the observer states". However, this assumption is not implicit in the mathematical requirements for a probability measure (additivity, summation to unity and non-negativity), and neither is it implicit in the definition of unitary wavefunction evolution, in which observers and measurements make no appearance. It is not even implicit in the idea of interpreting measurements in terms of bifurcating observers, which tells us only that observers split when making measurements, which could be compatible with any number of probability measures. Everett probably felt it was a necessary part of a sound interpretation of probability in terms of observers, but this clearly depends on one's interpretation of probability theory, which is a controversial topic, which I will explore in Ch. 4, and my own conclusions will not comply with Everett's assumptions of observer counting.

Everett's problem is clear: when it comes to the question of "what to count" when computing probabilities, he has made both *(1)* an assumption of amplitude dependence in stage 1, and *(2)* an assumption of observer counting in stage 2, neither of which he has justified. But since the first assumption asks us to count amplitudes, and the second to count worlds (or observers), won't these two assumptions yield conflicting probability measures? For Everett's scheme to work, it seems, the Born rule (which counts amplitudes) must be shown to yield the same results as counting observers, sequences or worlds.

Definition 3.11. We define two world-counting measures, E_n^W and W_n, exactly analogous to E_n^B and B_n, respectively, but based on a sequence-counting statistic. For our binary measurement example, this would be:

$$E_n^W(r_1 \cdots r_n) = \frac{\left| \frac{\sum_{i=1}^n r_i}{n} - \frac{1}{2} \right|}{\frac{1}{2}} \tag{3.40}$$

$$W_n(r_1 \cdots r_n) = 1 - E_n^W(r_1 \cdots r_n) \tag{3.41}$$

In the limit of infinite n, this yields exact statistical tests, B_∞ and W_∞, for whether a world is following Born or world-counting statistics, respectively:

$$B_\infty(r_1 \cdots) = \lim_{n\to\infty} B_n(r_1 \cdots r_n) \tag{3.42}$$

$$W_\infty(r_1 \cdots) = \lim_{n\to\infty} W_n(r_1 \cdots r_n) \tag{3.43}$$

Returning to our example, consider a world yielding exactly 3/4 results of $|1\rangle$, with the remainder of the results $|0\rangle$. For instance, the sequence $\{111011101110\cdots\}$. This gives:

$$B_\infty(11101110\cdots) = \lim_{n\to\infty} B_n(11101110\cdots r_n) = \lim_{n\to\infty}\left(1 - \frac{\left|\frac{\frac{3}{4}n}{n} - \frac{3}{4}\right|}{\frac{3}{4}}\right) = 100\% \tag{3.44}$$

$$W_\infty(11101110\cdots) = \lim_{n\to\infty} W_n(11101110\cdots r_n) == \lim_{n\to\infty}\left(1 - \frac{\left|\frac{\frac{3}{4}n}{n} - \frac{1}{2}\right|}{\frac{1}{2}}\right) = 50\% \tag{3.45}$$

Our test, then, tells us that this sequence follows the Born rule, but *not* world-counting.

Next, consider a world with half one result, and half the other. This is like flipping a coin indefinitely. We get:

$$B_\infty(101010\cdots) = \lim_{n\to\infty}\left(1 - \frac{\left|\frac{\frac{1}{2}n}{n} - \frac{3}{4}\right|}{\frac{3}{4}}\right) = 66.\bar{6}\% \tag{3.46}$$

$$W_\infty(101010\cdots) = \lim_{n\to\infty}\left(1 - \frac{\left|\frac{\frac{1}{2}n}{n} - \frac{1}{2}\right|}{\frac{1}{2}}\right) = 100\% \tag{3.47}$$

This world follows world-counting statistics, and *not* Born statistics.

Imagine a sequence with one $|1\rangle$, and the remainder $|0\rangle$s. This yields:

$$B_\infty(100\cdots000) = \lim_{n\to\infty}\left(1 - \frac{\left|\frac{1}{n} - \frac{1}{2}\right|}{\frac{1}{2}}\right) = 0\% \tag{3.48}$$

$$W_\infty(100\cdots000) = \lim_{n\to\infty}\left(1 - \frac{\left|\frac{1}{n} - \frac{3}{4}\right|}{\frac{3}{4}}\right) = 0\% \tag{3.49}$$

Not surprisingly, this world tests negative for *both* statistics.

It would seem likely, based on this, that for very long measurement sequences, and in the limit, we can expect Born worlds to be in the tiny minority of worlds. Everett wishes to show, rather, that in the infinite limit, the two statistics (amplitude-counting and world-counting) are, in fact, equivalent. He does this by making the following observation (paraphrased below):

In the infinite limit, the set of non-Born worlds (or sequences) has zero Born measure:

$$p^B_\infty(nonBorn_\infty) = 0\% \tag{3.50}$$

by which we shall mean

$$\lim_{n\to\infty,\,\epsilon\to 0} p^B_n(nonBorn_n(\epsilon)) = 0\% \tag{3.51}$$

where $nonBorn_n(\epsilon)$ is the set of all sequences of length n that are more than $\epsilon\%$ away from perfectly passing the Born test:

$$nonBorn_n(\epsilon) = \{r_1 \cdots r_n \,:\, B_n(r_1 \cdots r_n) < (1-\epsilon)\} \tag{3.52}$$

In other words, the Born measure of the set of all non-Born worlds approaches 0%, in the limit. Speaking informally, we may wish to describe $nonBorn_\infty$ as an infinite set of infinite-length non-Born sequences, with Born measure zero. It is worth noting that this is a *set* of zero measure, not a *sequence* of zero measure, since *all* infinite length individual sequences will have zero Born measure, even Born ones! But, even though each individual infinite Born sequence has measure zero ($p^B = 0$), the set of all infinite Born sequences clearly does not, since:

$$\lim_{n\to\infty,\,\epsilon\to 0} p^B_n(Born_n(\epsilon)) = 100\% \tag{3.53}$$

where

$$Born_n(\epsilon) = \{r_1 \cdots r_n \,:\, B_n(r_1 \cdots r_n) > (1-\epsilon)\} \tag{3.54}$$

The set of all Born sequences, then, forms a set of Born measure 1.

Now if a set of sequences has Born measure 1—or we could say 100%, since the Born measure is intended as a probability measure—then it exhausts all worlds represented in the wavefunction, since the wavefunction is a wave of amplitudes, and the Born measure is an amplitude-counting measure. Put another way, a sequence with zero measure also has zero amplitude:

$$p^B(r_1 \cdots r_n) = 0\% \rightarrow a_{1\cdots n} = 0 \tag{3.55}$$

since

$$p^B(r_1 \cdots r_n) = |a_{1\cdots n}|^2 \tag{3.56}$$

So here is the crux of Everett's phase 2 argument: if the non-Born worlds do not even appear in the wavefunction, then the size of the set $nonBorn_\infty$ is 0,

$$|nonBorn_\infty| = 0 \tag{3.57}$$

and these worlds must be excluded from the set of total worlds, even for the purposes of calculating p^w_∞ and W_∞, since none of these worlds are even represented in the wavefunction at all. Thus, all

the worlds represented in the wavefunction (in the infinite limit) are Born worlds. So even though our earlier arguments for finite length sequences/worlds legitimately showed that the Born measure and the sequence measure gave different results, these results necessarily converge in the limit, given that the non-Born worlds converge to zero amplitude. Hence, the two statistics are equivalent in the limit. Q.E.D. (according to Everett). □

3.3.5 The Limits of World-counting

What Everett believes he has shown is that, in spite of the fact that non-Born-worlds for finite n outnumber the Born worlds, that in the infinite limit, the non-Born worlds disappear from the picture, melting away to measure zero, and so amplitude-counting (*i.e.*, the Born rule) has been shown to be equivalent (in the limit) to observer-counting (which Everett has assumed as an *a priori* standard).

There is a general consensus, however, that Everett's proof does not follow through. The key move in Everett's original proof is not given formal expression, but is expressed rather in purely verbal terms. If we try to formalize it in terms of limits—rather than informal, questionable talk about completed infinities—we find that Everett's point is not nearly so convincing.

Recall, that the key point was that

$$|nonBorn_{\infty}| = 0 \tag{3.58}$$

Or, more formally,

$$\lim_{n\to\infty,\,\epsilon\to 0} |nonBorn_n(\epsilon)| = 0 \tag{3.59}$$

Let's distinguish now between $nonBorn_n(\epsilon)$, which does not include zero amplitude worlds, and $nonBorn_n^{possible}(\epsilon)$, which includes all the possibilities, regardless of amplitudes (this, then, was our original conception of the number of non-Born worlds, before Everett's correction). What is being claimed here by Everett is that this distinction allows us to modify the total number of worlds when doing world-counting (in the limit). Recall how we calculate the sequence-probability p^w, of a world:

$$p_n^w(r_1 \cdots r_n) = \frac{1}{b^n} \tag{3.60}$$

Everett is claiming we can omit certain worlds from consideration, and these are the worlds with exactly zero (not just small) amplitudes. So we modify the above expression to exclude the zero Born-measure worlds:

$$p_n^w(\{r_1 \cdots r_n\}) = \sum_{r_1 \ldots r_n} \frac{1}{b^n - |\{r'_1 \cdots r'_n \,:\, p_n^B(r'_1 \cdots r'_n) = 0\}|} \tag{3.61}$$

If we take a straightforward infinite limit of this, however, for some set of worlds, $\{r_1 \cdots r_n\}$ that we have defined in such a way that it can be taken to the limit of infinite n:

$$\lim_{n\to\infty} p_n^w(\{r_1 \cdots r_n\}) = \lim_{n\to\infty} \sum_{r_1 \ldots r_n} \frac{1}{b^n - |\{r'_1 \cdots r'_n \,:\, p_n^B(r'_1 \cdots r'_n) = 0\}|} \tag{3.62}$$

we hit a problem for Everett. The problem is that there *are no worlds in the power set of sequences that have zero Born measure.* We may have spoken informally about an infinite sequence that had measure zero, but now we are trying to construct the limit that will give rise to this conclusion, in terms of finite n, and there will always be some (perhaps tiny) amplitude for *any* sequence. So:

$$\forall n \left|\left\{r_1 \cdots r_n \,:\, p_n^B(r_1 \cdots r_n) = 0\right\}\right| = 0 \tag{3.63}$$

and

$$\begin{gathered}\lim_{n\to\infty} p_n^w(\{r_1 \cdots r_n\}) = \\ \lim_{n\to\infty} \sum_{r_1 \ldots r_n} \frac{1}{b^n - |\{r'_1 \cdots r'_n \,:\, p_n^B(r'_1 \cdots r'_n) = 0\}|} = \\ \lim_{n\to\infty} \sum_{r_1 \ldots r_n} \frac{1}{b^n} = \\ \lim_{n\to\infty} \frac{|\{r_1 \ldots r_n\}|}{b^n}\end{gathered} \tag{3.64}$$

which is exactly what we would get with our original sequence measure. Nothing has been omitted from the wavefunction because an infinite limit has to be expressed as a limit of a finite expression, and for finite n, there is no disappearance of amplitude.

We might try to save Everett here by constructing our limit differently. Perhaps we will try first to find the limit of the number of worlds, which will be less than the power set, and then plug *this* into our limit expression. But this does not help, for it is, alas, still not less than the power set, since, once again, there are no worlds at each finite n that have zero amplitude:

$$\begin{gathered}\lim_{n\to\infty} (\#\text{ of worlds}) = \\ \lim_{n\to\infty} \left(b^n - \left|\left\{r'_1 \cdots r'_n \,:\, p_n^B(r'_1 \cdots r'_n) = 0\right\}\right|\right) = \\ \lim_{n\to\infty} b^n = \infty\end{gathered} \tag{3.65}$$

In any case, finding the limit of the number of worlds is not valid to begin with. You cannot simply replace a term in the denominator of a fraction with its limit, and then plug that into the overall expression.

We could perhaps try to introduce a tolerance, ϵ, in hopes that then we can get a nonzero set of worlds to omit for finite n. But this doesn't work either. Imagine we have defined some means of

measuring p^B to within some tolerance level ϵ, which we call $p^{B,\epsilon}$. This will give us

$$\lim_{n\to\infty} p_n^w(\{r_1 \cdots r_n\}) = \lim_{n\to\infty,\,\epsilon\to 0} \sum_{r_1 \ldots r_n} \frac{1}{b^n - \left|\left\{r'_1 \cdots r'_n \,:\, p_n^{B,\epsilon}(r'_1 \cdots r'_n) = 0\right\}\right|} \tag{3.66}$$

Now we note that

$$\left|\left\{r'_1 \cdots r'_n \,:\, p_n^{B,\epsilon}(r'_1 \cdots r'_n) = 0\right\}\right| \neq 0 \tag{3.67}$$

and so (we hope) we can legitimately eliminate worlds in the limit. However, this does not work, since the introduction of ϵ means that we are not actually counting zero amplitude worlds here, but very tiny-amplitude worlds. These worlds are still represented in the wavefunction, and so our justification for eliminating them in the first place is no longer there. We cannot eliminate zero-amplitude-to-some-tolerance worlds from the wavefunction. The whole nature of the argument is that these worlds can only be eliminated because *they do not exist at all*, not because they exist and have some tiny amplitude. Recall that we are trying to show that amplitude-counting amounts to the same thing as world-counting. Thus, eliminating worlds because their amplitudes are tiny is a circular argument. We would be saying that amplitude-counting amounts to world-counting, if only we agree to count worlds according to their amplitudes.

In summary, the basic problem with Everett's stage 2 is that it depends on an assumption that I will call

Assumption 3.12. *"Amplitude-attenuation of the world-count": if the amplitude of a world (a sequence of length n) goes to zero in the limit (as n goes to infinity), then the contribution made to the world-count also goes to zero.*

This assumption can go completely unnoticed if we fail to recognize that there is a difference between the *amplitude* of a sequence, and its contribution to the world count.

Definition 3.13. Define a function $W()$ that returns a sequence's contribution to the world count, given a particular observation sequence:

$$W(S_k^n) = \begin{cases} 1 & \text{if amplitude } \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.68}$$

where

$$S_k^n = \text{the } k\text{th sequence of length } n$$

The world-count contribution is simply 1 for any world that is actually represented in the wavefunction, regardless of amplitude (zero amplitude worlds are eliminated, not because of their particular amplitude value, but because, being zero amplitude, they are not represented in the wavefunction).

Probability can be defined in terms of world counts as follows:

$$p(S_k^n) = p(S^n) = \frac{1}{\sum_k W(S_k^n)} \tag{3.69}$$

In other words, all worlds are counted equally (this is the whole point of world-counting as an *a priori* measure: if all worlds are equally real, they must be equally probable).

Now, if S_i^∞ (the ith infinite sequence/world) is a non-Born-rule infinite sequence, then if we consider the first n elements, as n gets larger and larger, we see that no matter how large n gets, even though the sequence's amplitude is getting smaller and smaller, approaching 0 in the limit, that the world count for the sequence does *not* approach zero. In fact, it remains at 1:

$$\lim_{n\to\infty} \sum_{k=1}^{m} W(S_k^n) = m \tag{3.70}$$

in contradiction of the assumption of amplitude-attenuated world-counting. The problem here is that the $W()$ function that is actually used to count worlds is not continuous, but has a sudden discontinuity at zero amplitude. Even at 0.00000001 amplitude, for instance, the count is still 1. It does not hit 0 until the amplitude is *exactly and absolutely zero*. Thus, while it may be true that the *amplitude* approaches zero in the limit, the amplitude is *not* what is being taken in the limit, the *world count* is. It may seem intuitive to say that because the amplitude in the limit is zero, then the world count is also zero, since the world count for a zero amplitude sequence would be zero, but this is actually not mathematically sound, since it takes an infinite limit as if it were a completed infinity, and plugs the result into another function. This is not how limits work.

This is not something unusual or esoteric that arises in this situation alone; it is a very familiar and common scenario for functions of x that divide by x, and are thereby discontinuous at $x = 0$. To take a simple example, unrelated to physics, assume that:

$$f(x) = \frac{x}{x-1} \tag{3.71}$$

$$g(x) = \begin{cases} 1 & \text{if } x \neq 1 \\ 2 & \text{otherwise} \end{cases} \tag{3.72}$$

Note that, at $x = 1$, $f(x)$ is undefined and $g(x)$ is discontinuous.

Then:

$$\lim_{x\to\infty} f(x) = 1 \tag{3.73}$$

$$g(1) = 2 \tag{3.74}$$

But note that:

$$\lim_{x\to\infty} g(f(x)) = \lim_{x\to\infty} \begin{cases} 1 & \text{if } \frac{x}{x-1} \neq 1 \\ 2 & \text{otherwise} \end{cases} = 1 \tag{3.75}$$

But if we followed Everett's use of the limit for world-counting, we could also get:

$$\lim_{x\to\infty} g\left(f(x)\right) = g\left(\lim_{x\to\infty} f(x)\right) = g(1) = 2 \tag{3.76}$$

And thus if the method of Everett's proof (stage 2) is valid, we can derive:

$$1 = 2 \tag{3.77}$$

This contradiction was derived because we plugged a completed infinity into a function as if it were a number. This is just what Everett's stage 2 proof is guilty of. It assumes that because the amplitude approaches 0 in the limit, and the count for a sequence is 0 if the amplitude is 0, that the count must also approach 0. But this does not follow, because the count is a discontinuous function at 0, and it actually approaches 1 in the limit, not 0.

In light of this, the only way to get the proof to work is to somehow cut small-amplitude worlds down to zero amplitude for large n. But the only way to do this is to assume that *somehow* amplitudes "count"... that the smallness of a world's amplitude can somehow matter to its *W()* value. Unfortunately, there is no way to do that without falling into circularity, since it is the relevance of amplitudes to the probability measure that the objectors are asking us to prove in the first place. Indeed, the most common critique of Everett's stage 2 proof—as well as other allied frequentist Born rule proofs—is that they are circular in nature, and in some way or another, assume what they are trying to prove (amplitude-counting) [46, 125].

It is also argued that, even if we grant the validity of Everett's infinite limit, the whole idea displays a problem for very large (as opposed to infinite) sequences. After all, if a result about infinite sequences is meaningful, we should be able to posit extremely large sequences that display properties close to the infinite limit (since there are, after all, no literally infinite sequences in the world). As it turns out, however, in such cases, we have more worlds that do *not* follow the Born rule than worlds that do (even assuming the number that do not goes to zero in the limit) . These non-Born-rule worlds are sometimes called "maverick worlds". They are (unsurprisingly) of low amplitude, but we cannot thereby throw them out without falling into the circularity objection, since the rejection of low-amplitude worlds is, after all, part of what we are trying to prove in the first place. Thus, the circularity objection to Everett's proof would remain, even if we were to grant the validity of his limit.

Thus, whatever the validity of Everett's use of limits, it seems undeniable that in the case of finite, large n, there are undeniably many more "maverick" worlds that disobey the Born rule than

those that follow it. Therefore, by Everett's own assumptions, we should *not* observe the Born rule. I will argue that this objection to Everett is sound, since Everett does seem to assume *a priori* relevance to world-counting, else he would not be trying to prove that amplitude-counting is equivalent to it in the limit. I will argue, however, that in order to be true to most of the rest of his principles, Everett should have simply ignored world-counting as an unjustified assumption.

If Everett had done this—by simply throwing out stage 2 as a misguided attempt to prove something that does not need proving—he would still have been open to the Born rule objection, however, since his stage 1 is based on amplitude dependence. Hence, the objectors would still have asked, "Why count amplitudes?" While he might have responded to this with a bemused "What else?", the objectors would certainly have asked for more than that. However, what the objectors can no longer do, I hope, is simply trot out world counting as the failed *a priori* standard that Everett should be accountable to, since world-counting is no more *a priori* justified than counting amplitudes.

3.3.6 The Hartle Proof

Hartle's proof [103] is usually considered to be of the same general character as Everett's, and they are frequently lumped together as being essentially the same. However, while Hartle's proof does appear superficially to be a more formal version of Everett's, he does *not* present the proof as being within an MWI perspective, nor does he present it as an attempt to derive amplitude-counting from world-counting. Since he does not interpret his infinite limit expression as being about world counts, his result does not suffer from the problems of Everett (but nor does it attempt to be as far-reaching).

The Hartle proof still suffers from the same problem as stage 1 of Everett's proof—it assumes that amplitudes are what matter. It essentially rolls this assumption right into something like Everett's stage 2, rather than presenting them as separate proofs. Hartle constructs an operator—the frequency operator—that calculates the relative frequencies of outcomes. Hartle then shows that this operator converges on the Born rule operator in the limit of an infinite number of repeated measurements. (If one follows a frequentist account of probability, then this frequency operator *is* a measure of probability.)

Suppose we have a system consisting of n identically prepared subsystems:

$$|\Psi_n\rangle = |\psi_1\rangle\,|\psi_2\rangle\,|\psi_3\rangle\cdots|\psi_n\rangle \tag{3.78}$$

We can abbreviate the above to

$$|\Psi_n\rangle = |\psi_1\psi_2\psi_3\cdots\psi_n\rangle \tag{3.79}$$

We will consider an operator $\hat{O}$ corresponding to an observable $\mathcal{O}$, and an eigenbasis $\{|i\rangle\}$ which diagonalizes $\hat{O}$:

$$\hat{O}\,|k\rangle = o_k\,|k\rangle \tag{3.80}$$

Next, we define a set F of frequency operators for measurement result x, $F = \left\{\hat{F}_n(x)\right\}$, such that, for a state $|\Psi_n\rangle$,

$$\hat{F}_n(x)\,|\Psi_n\rangle = f(x,\Psi_n)\,|\Psi_n\rangle \tag{3.81}$$

where $f(x,\Psi_n)$ is the frequency of the individual measurement result o_x in $|\Psi_n\rangle$. Note here that each of the identical subsystems has had only a *single* measurement taken. The frequency represents how often the particular eigenvalue x of $\hat{O}$ results from applying $\hat{O}$ once to n identical subsystems. The eigenstates of $\hat{F}_n$ are ensembles of identically prepared substates, and their eigenvalues are obtained from sets of eigenvalues of $\hat{O}$, in which we can count out how many are equal to x.

The frequency operator is defined more precisely in [103] as:

$$\hat{F}_n(x) = \sum_{k_1,\ldots,k_n} f\left(x,|k_1\cdots k_n\rangle\right)|k_1\cdots k_n\rangle\langle k_n\cdots k_1| \tag{3.82}$$

where

$$f\left(x,|k_1\cdots k_n\rangle\right) = \frac{\sum_{i=1}^{n}\delta_{xk_i}}{n}$$

This is a sum of projectors onto all the possible eigenstates, $|k_1\rangle|k_2\rangle|k_3\rangle\cdots|k_n\rangle$, that could result from the measurement of $\hat{O}$ on each of the members of the ensemble. These are weighted by the fraction of k_i that are x. Each such eigenstate will be associated with a sequence of eigenvalues, representing the outcomes of the $\mathcal{O}$ measurements, $\{x_1,x_2,\ldots,x_n\}$. While these sequences can be thought of as resulting from simultaneous measurements on members of an ensemble, it can also be intuitive to imagine the members as being prepared and measured sequentially in time, so that the ensemble can be thought of similarly to a sequence of coin-flips, where each possible $\{x_1,x_2,\ldots,x_n\}$ is thus a different possible world history.

Applying this operator to any one particular eigenstate of the ensemble gives us

$$\hat{F}_n(x)\,|\psi_1\cdots\psi_n\rangle = \sum_{k_1,\ldots,k_n} f\left(x,|k_1\cdots k_n\rangle\right)|k_1\cdots k_n\rangle\langle k_n\cdots k_1|\,\psi_1\cdots\psi_n\rangle \tag{3.83}$$

Note that other than $f()$, this is just application of expansion of the identity operator:

$$\hat{I} = \sum_{k_1,\ldots,k_n} |k_1\cdots k_n\rangle\langle k_n\cdots k_1| \tag{3.84}$$

So what we are doing here is breaking $|\psi_1\cdots\psi_n\rangle$ into a sum of components, each of which is weighted by its amplitude, $\langle k_n\cdots k_1|\,\psi_1\cdots\psi_n\rangle$, which could be zero, and corresponds to one possible

measurement sequence $\{x_1, x_2, \ldots, x_n\}$:

$$|\psi_1 \cdots \psi_n\rangle = \hat{I} |\psi_1 \cdots \psi_n\rangle = \sum_{k_1,\ldots,k_n} |k_1 \cdots k_n\rangle \langle k_n \cdots k_1| \psi_1 \cdots \psi_n\rangle \tag{3.85}$$

Applying the frequency operator to the ensemble state has the effect of weighting each of its components by how often x occurs in it:

$$\hat{F}_n(x) |\psi_1 \cdots \psi_n\rangle = \sum_{k_1,\ldots,k_n} f(x, |k_1 \cdots k_n\rangle) |k_1 \cdots k_n\rangle \langle k_n \cdots k_1| \psi_1 \cdots \psi_n\rangle \tag{3.86}$$

Note that each possible sequence, or post-measurement "world", is weighted *both* by

1. its amplitude in the wavefunction, $\langle k_n \cdots k_1| \psi_1 \cdots \psi_n\rangle$, and
2. the frequency of $|x\rangle$ in it, $f(x, |k_1 \cdots k_n\rangle)$.

This is an important point, because this is exactly the place where this derivation will be attacked by the critics, who will say that if this is truly a derivation from the wave-function alone with no collapse, then it should not assume that probability is tied to amplitude (meaning higher amplitude always yields higher probability). The Born rule critics might (perhaps) insist that worlds should only be weighted by #2 above:

$$\hat{F}_n(x) |\psi_1 \cdots \psi_n\rangle = \sum_{k_1,\ldots,k_n} f(x, |k_1 \cdots k_n\rangle) |k_1 \cdots k_n\rangle \quad ? \tag{3.87}$$

Keeping this point of potential controversy in mind, let us continue with the Hartle proof. Frequentists generally take probability to be relative frequency in the infinite limit, so next, we need to extend $F()$ so that it can be applied to an infinite ensemble, $|\psi_1 \cdots\rangle$:

$$\hat{F}_n(x) |\psi_1 \cdots\rangle = \hat{F}_n(x) |\psi_1 \cdots \psi_n\rangle \otimes |\psi_{n+1} \cdots\rangle \tag{3.88}$$

In other words, we apply our finitistic frequency operator $F_n()$ to an infinite ensemble by applying it to the first n members of the ensemble, and leaving the rest alone. Hartle then proves the Born rule, in the limit of infinite sequences, by showing that the frequency in the limit is the same as the norm-squared amplitude, or more precisely, that the norm-squared amplitude acts, in the limit of infinite n, as an eigenvalue of the frequency operator, with an ensemble of n identical subsystems as the eigenvector:

$$\lim_{n\to\infty} \hat{F}_n(x) |\psi_1 \cdots \psi_n\rangle = \lim_{n\to\infty} |\langle x|\psi\rangle|^2 |\psi_1 \cdots \psi_n\rangle \tag{3.89}$$

$$= |\langle k_n \cdots k_1| \psi_1 \cdots \psi_n\rangle|^2 \tag{3.90}$$

I will not go through all the algebra here, but it is useful to note that they key facts that Hartle makes use of are normality and additivity, as in Everett's stage 1. In fact, Hartle's proof is a kind

of combination of Everett stage 1 and 2 rolled into a single cohesive proof. Like Everett's stage 1, Hartle's proof assumes amplitude dependence, and so is circular if used as a true Born rule proof in the sense demanded by the objectors. Like Everett's stage 2, Hartle's proof is essentially examining a world history (not a single branch from one observation) and asking what happens in the limit.

One would think, then, that it suffers from the same problem of limits as Everett's stage 2. But *technically* it doesn't, because Hartle never claims that he is proving that all the histories are equally probable. He simply shows that the frequency operator approaches the Born rule in the limit, which it does. He is thus not guilty of assuming, implicitly or otherwise, that world-counting is the *a priori* method of counting, as Everett was. We can't use, against Hartle, the argument we used against Everett, that he is discounting small amplitude worlds because they disappear in the limit, even though they outnumber other worlds for large n. Hartle is not claiming that the small amplitude worlds disappear in the limit. He does, however, assume (implicitly in Eqn 3.87), that amplitudes are the countable entity for frequency counts, and then proves that such a frequency count produces the Born rule, in the limit.

Hartle's proof is thus more mathematically correct than Everett's, but also not nearly as ambitious. In a sense, it still fails, but not due to incorrectness, but simply due to lack of relevance. The fact that the frequency operator approaches the Born rule in the limit is not relevant to the Born rule objectors, since, in the limit, the world-counts are still at complete variance with the Born rule. Hartle has produced an error-free, correct, but not especially relevant version of Everett's proof.

Because there is still the assumption of amplitude dependence, the Born rule objectors will not be happy with Hartle's proof, anyway. However, what if we assume that the "what-else?" argument has been successful, and we have decided that there is no other elegant, analytic countable in the wavefunction, other than amplitudes, and so we should accept them as the countable ontic entities of the wavefunction? In this case, however, Hartle's proof still does not fundamentally go much beyond Everett's stage 1. Equation (3.87) is more or less equivalent to Everett's assumption of amplitude dependence. Once we accept that, Everett's stage 1 proof should already be sufficient to tell us that the relative frequency operator will approach the Born rule in the limit, since it tells us that the norm-squared is the only possible measure. Everett's stage 2 was intended for an entirely different purpose, in that it also purports to show that the limit-based frequency operator is essentially counting worlds or histories in the long-run. Hartle, presumably, could respond to this by contending that there is no need for world histories to be equally probable, since the frequency operator converges on the Born rule due to the amplitude-counting assumption, *not* due to the dwindling of the number of maverick worlds in the limit. Of course, since Hartle's amplitude-counting assumption is merely implicit, and not explicitly stated as in Everett stage 1, one could posit either:

1. that there was no intent to assume amplitude dependence, and the proof was intended to be just as strong as Everett intended his stage 2 to be, in which case, the proof fails because it actually *does* assume amplitude dependence, or

2. that the assumption of amplitude dependence was intentional, in which case, the proof is correct, but not as strong as Everett's stage 2 is intended to be (being more like a straightforward extension of stage 1 to infinite sequences).

Either way Hartle's proof is not a workable response to the Born rule objectors, whether it was intended as such, or not.

3.3.7 The Farhi-Goldstone-Gutmann Proof

The Farhi-Goldstone-Gutmann proof [83] attempts to fix the main problems with Everett's and Hartle's proofs, and there are some who believe that it does, and is the final word on the subject. But even if it does, it does so only by invoking completed infinities, something that is not consistent with the algorithmic epistemology I will be assuming in this dissertation. However, even aside from my own philosophical problems with completed infinities, there are reasons to be especially suspicious of such an approach, in this case.

The Farhi-Goldstone-Gutmann proof can be viewed as essentially an attempt to generalize Hartle's result:

$$\lim_{n\to\infty} \hat{F}_n(x) \left|\psi_1 \cdots \psi_n\right\rangle \quad = \quad \left|\left\langle k_n \cdots k_1\right| \psi_1 \cdots \psi_n\right\rangle\right|^2 \tag{3.91}$$

from a statement about the behaviour of a *finite* frequency operator on n identical systems, to one about an *infinite* frequency operator on an infinity of identical systems. The proof essentially purports to derive an infinitistic generalization of Hartle's result, which could be expressed as:

$$\hat{F}_\infty(x) \left|\psi_1 \cdots\right\rangle \quad = \quad \left|\left\langle \cdots k_1\right| \psi_1 \cdots\right\rangle\right|^2 \tag{3.92}$$

Now if one *does* allow the fiction of talking about probabilities on *completed* infinite ensembles (meaning outside of the context of infinite limits), then it is easy to see that this infinitistic Born rule proof *should* be expected to carry through. Recall that, *in the limit*, the amplitude of non-Born rule sequences approaches zero (even though in the limit they outnumber the Born rule sequences). However, if we drop the idea that we are dealing with limits, and imagine that we actually have an infinite sequence, and can usefully talk about *its* probability, then we are allowing ourselves to make the leap of faith that this sequence actually has an *actual* amplitude of exactly zero. Hence, the probability of the non-Born rule sequences *is* zero. In fact, there *are* no non-Born rule sequences. In fact, we don't even need to sneak in amplitude dependence anymore, since these zero-amplitude

worlds do not exist at all in this infinitistic wavefunction, automatically guaranteeing "amplitude dependence".

I will leave my specific beefs with infinitistic interpretations of probability theory to Ch. 4 (but the reader may guess that I do not find them to be especially coherent). I will say here only that, even if one accepts this kind of argument—perhaps from the perspective of an infinitistic philosophy of mathematics—there remains the problem that, for any very large finite sequence (no matter how large), the number of non-Born worlds will seriously outnumber the Born worlds. Since any actual sequence of measurements is finite, it still seems that the proof fails in the real world.

3.3.8 Hanson's Mangled Worlds

In Hanson's system [102], "inexact decoherence can allow large [amplitude] worlds to drive the evolution of very small [amplitude] worlds, 'mangling' those worlds. Observers in mangled worlds may fail to exist, or may remember events from larger worlds." This is, in my opinion, an extremely important general idea, quite aside from the specific use that Hanson puts it to.

Hanson first assumes the standard model of environmentally-induced decoherence, in which off-diagonal terms in the density matrix represent correlation with an environment, and are extremely tiny compared to the diagonal elements, yielding an almost-diagonal matrix. He points out, however, that it is quite plausible that the so-called "tiny" off-diagonal terms for a large-amplitude world will still be quite huge compared to the main diagonals for a small-amplitude world. If so, then the overall dynamics of the small-amplitude worlds may actually be driven by the dynamics of the larger-amplitude worlds. In a sense, the assumption that an almost-diagonal density matrix destroys interference with other worlds is incorrect for an extremely low-amplitude world. It fails, and it fails for synthetic *a priori* reasons: it is not possible to imagine a coherent and consistent history and memory state for an observer in such a world. In other words, the small amplitude worlds, while they "exist"—in the sense of being in the wavefunction with nonzero amplitude—are not stable and cannot support observers.

Hanson uses this idea to argue that it might be possible to prove the Born rule via world counting once world mangling is taken into account. I am not so impressed with this use of it, since I am not convinced that world-counting is a valid *a priori* in the first place. However, the general idea can be used all on its own—without attempting to prove the Born rule—as a disproof of *a priori* world-counting. After all, if some small-amplitude worlds that are analytically there in the wavefunction actually fail to exist because it is not possible for an observer to have a stable history of memories laid down in it, then whatever the ultimate method of calculating probabilities, it is *not* a straightforward count of worlds, at least not in the traditional sense. Not only that, it

is the small amplitudes of these worlds that makes them susceptible to mangling, so we can also say that not only is world-counting not a straightforward process of counting worlds, but also that, for the same reasons, amplitude matters. Not only does it matter, but it is *larger* amplitudes that count more—at least in some situations—than smaller amplitudes. The mere fact that it is *possible* to imagine a situation where this is so disproves straightforward world-counting as the method of probability calculation, and opens the door for some degree of amplitude relevance (although not necessarily full amplitude dependence, in the sense I have been using that term).

Hanson's goals are somewhat different from mine. He is seeking to derive the Born rule from world-counting, in the tradition of Everett, using world-mangling as his means of "weeding out" small-amplitude worlds. However, he needs to assume certain things about the statistics of small-amplitude worlds in order to get the Born rule to emerge, and these are worth listing here.

Assumption 3.14. ***Hanson's Conjecture #1:*** *"After two worlds split due to a decoherence event, their coherence $\epsilon(t)$ typically falls with time t, but eventually falls slower than $e^{\sqrt{rt}}$, where r is the effective rate of further decoherence events."* [102, p.1137]

Hanson notes that—as we already discussed in §2.5.4—environmentally-induced decoherence, in many situations, occurs exponentially fast (recall that this is the basis of the conclusion that even light scattering off the outside of Schrödinger's cat box will cause an effective collapse to either the dead or alive state, relative to the human observer, within a very short span of time). However, Hanson notes that this exponential decay in the off-diagonals does not typically approach an asymptote of zero, but rather some very tiny but non-zero value. So long as the decoherence is thus kept in check, and the tiny off-diagonals do not become too tiny too quickly, then the dynamics of the small-amplitude worlds will be largely driven by the large-amplitude worlds. In other words, such small-amplitude worlds have failed to actually "split" at all.

This is inherently a synthetic *a priori* and not analytic feature of the wavefunction—but this is to be expected, since the whole notion of a world or observer within the wavefunction is synthetic. So Hanson is not saying that small-amplitude worlds cannot be analytically specified. However, he is saying that observers in such worlds will fail to stably *inhabit* that world, in any useful sense. Even if we conjecture that they continue to exist at all, their memories will be consistent with a larger-amplitude world, refuting the idea that they were ever really "in" the small-amplitude world in the first place. Hence, Hanson's second conjecture.

Assumption 3.15. ***Hanson's Conjecture #2:*** *"When the coherence ϵ between two worlds is large enough compared to their relative measure δ, human observers in the small world will typically be 'mangled,' i.e., will either fail to exist or will remember the measurement frequency of the large*

world." [102, p.1138]

If we assume—as I do—that "worlds" are defined entirely *a priori* synthetically in the first place, then there is no consistent way to define such small-amplitude worlds as existing in the first place. They are not "worlds", since they do not permit stable observers to lay down consistent memory records within that world.

This is all very similar to my own independent (but later) model [174] of maverick worlds failing to exist due to information-theoretic instability (which will be discussed in Ch. 6). Hanson's idea is developed from a different perspective, and I have not in the past used the "mangled world" metaphor to describe it; however, I think it is essentially the same result, so in this dissertation I will sometimes adopt Hanson's terminology for the general idea.

The basic idea of world-mangling holds independently of whether we attempt to use it to prove the Born rule. It explains why maverick "worlds" may not be countable as "worlds" at all. I will suggest not even calling such worlds "physical", but leaving them in the category of "logically possible".

Hanson wishes to go further, however, than merely showing that maverick worlds fail to exist. By thus discounting these small-amplitude worlds, he wishes to weed them out of the world-count in an Everett-style frequentist proof of the Born rule. This sounds promising, since discounting small-amplitude worlds does seem to be what is required for such a proof. However, it still does not follow straightforwardly from world-mangling (via Hanson's first two conjectures) that the long-run statistics of measurements will match the Born rule. But Hanson notes that in typical measurement situations, where we test the Born rule, there are actually a large number of uncounted background decoherence events happening. Hanson attempts to quantify what would be required of the statistics of such events in order to recover the Born rule (or in order to come very close to recovering it), leading to his third conjecture.

Assumption 3.16. ***Hanson's Conjecture #3:*** *"Typical situations where we test the Born rule are the result of a large number ($> 10^4$) of mostly uncounted decoherence events, each of which has a small fractional influence. Even when we only count frequencies from a few events, many other background events occur. Thus the distribution of world size is lognormal, with [standard deviation] σ large (> 50)."* [102, p.1139]

Hanson shows that if this conjecture holds, then for any two randomly chosen worlds, one will eventually mangle the other, implying that only worlds above a certain amplitude threshold will survive in the long-run, and that smaller than normal worlds will almost surely be mangled. There is thus a threshold or "transition region" as we go from smaller to larger-amplitude worlds, above which worlds survive, and below which they eventually get mangled. Hanson derives the required

constraint on the placement of this threshold in order to recover the Born rule by a frequentist argument. This is his fourth conjecture.

Assumption 3.17. ***Hanson's Conjecture #4:*** *"... There is an outcome independent transition region $[\underline{m}, \overline{m}]$ in world size m, below which worlds are mangled and above which they are not. For all $m \in [\underline{m}, \overline{m}]$, we have $\left|\log \frac{m}{\hat{m}}\right| \ll \sigma^2 \gg 1$."* [102, p.1141]

While the basic idea of world-mangling, via the first two conjectures, will be echoed later in Ch. 6, in my system it will not be necessary to employ the other conjectures to derive a frequentist account of the Born rule in terms of world-counting, since I will not be accepting the frequentist view of probability or the assumption of branch-counting, which are required to motivate such a project. Under my assumptions, we are looking instead to justify the Born rule without appeal to long-run frequencies, appealing only to the structure of single cases.

However, even if we do not accept world mangling as a solution to the Born rule objection, it remains (via the first two conjectures) an invaluable tool for understanding how maverick worlds can be, not only improbable, but actually impossible—even if they do not disappear to zero amplitude. I will argue in Ch. 6 for a notion of cosmic thermodynamic stability, based on very similar ideas, which could potentially explain the very physicality of the external world.

3.3.9 Buniy's Discrete State Space Solution

In [32], small-amplitude worlds are discarded, and the Born rule thereby recovered, by making the state-space discrete at a very small scale. This allows very small amplitude sequences to hit zero. This, of course, requires its own assumption: the discrete nature of the state space. The authors argue that, since it cannot be verified whether the state space is discrete or not, given that any measurements will always be of finite precision, one could argue that the current physical theory ought to be agnostic as to whether the state space is discrete (although it is granted that physicists tend to make the assumption of a continuum, the authors rightly point out that this is nothing more than an unjustified bias).

While there is some attractiveness to this point of view, there can be little doubt that it does not prove the Born rule from the wave function mechanics. Like Hanson's version, it attempts to give a rationale for why amplitudes might be cut off at some threshold, and it likewise presumes—along with all the other frequentists proofs—that world-counting provides some kind of *a priori* standard that we ought to be shooting for.

3.3.10 The Deutsch-Wallace Proof

3.3.10.1 Introduction

Bayesians will tend to disapprove of all the Everett-style frequentist proofs, quite aside from the issue of their correctness as proofs, and largely on the basis that they *are* frequentist, since Bayesians view frequentism as an out-dated and inadequate formulation of probability. The Deutsch-Wallace Born rule proof [71, 225, 227] seeks to address this by deriving the Born rule in broadly Bayesian terms, and in particular, follows the tradition of Savage's decision theory [186]. The probability measure is derived within the context of a rational agent performing acts, for which he is rewarded or not rewarded, in an Everettian universe where he *knows* the universe is Everettian (and also knows the wavefunction state to whatever degree of detail is relevant to his situation).

Wallace, himself, does not take a hard-core (subjectivist) Bayesian stance, accepting at least the possibility that there are objective probabilities [227, p 231], and Deutsch—at least according to Wallace [225, p 316]—has serious reservations about the whole subjectivist conception of probability. In addition, it is arguable (as we will soon see) that some of Wallace's assumptions are decidedly objectivist in nature. Thus, we might say that the Deutsch-Wallace proof is a "subjectivist derivation of objective probabilities", since the hypothetical rational agent in the proof knows everything (or at least everything that is relevant). As a result, there is no uncertainty in the agent's mind about reality. For hard-core Bayesian subjectivists, this already might present a problem, since probability for them is literally supposed to *be* a measure of uncertainty, leading to the question of how there can be probabilities (other than 0 and 1) under such circumstances. For a probability *objectivist*, however, this does not automatically create a problem, since he views probabilities as having objective content independent of any observer, and hence we do not need to always relate them to degrees of uncertainty. Wallace discusses proposed solutions to this problem, and throws his hat into the ring of "subjective uncertainty", in which there is a kind of uncertainty that can subjectively exist for an agent who nonetheless has complete knowledge, and hence absolutely no *objective* uncertainty. This "SU" solution, as Wallace calls it, is essentially sound, in my opinion. However, I will not use this language, myself, as I believe that it leads to great confusion, the choice of the words "subjective" and "objective" being, in this context, poorly chosen (for reasons I have explained at length in Ch. 1). When quantum probabilities are more properly framed as "synthetic *a priori*", rather than "subjective", I believe the argument against the SU camp loses its bite (we will see in more detail why this is so, in coming chapters).

Once we accept a synthetic (but still *objective*) notion of quantum probability, this brings into question the very use of a subjectivist framework to begin with. If we are *assuming* that our agent

knows everything there is to know, why can we not drop the subjectivist pretense, and directly employ an objectivist interpretation of probability, to derive our probability rule more directly? Why is the decision-theoretic framework needed? Of course, if you are a hard-core subjectivist Bayesian, the answer is that you believe that probabilities *are* subjective, by definition. Any talk about probabilities being "objective" then, has to be justified as a special case, in terms of subjective probabilities. We will explore these issues in more detail in the next chapter, on the interpretation of probability theory. For now, suffice it to say that I personally find the approach of deriving objective probabilities within a subjectivist framework more than a little awkward and unnecessary. I will be throwing my own hat into the ring of the probability debate in the next chapter, but for the time being, I will try to show that the meat of Wallace's "proof" is not even in the decision-theoretic part, nor in the formal proof part, but rather in the proof's axioms, for which Wallace gives purely verbal and informal arguments (not to disparage verbal arguments, but it gives us pause to wonder whether calling the result a "formal Born rule proof" is exactly appropriate).

It will not be necessary here to examine Wallace's proof in full detail. I will explain the overall gist of the proof, and examine details only for those aspects that may be questionable. Much of the exposition that follows is taken directly from [227]; however, I have changed some of the terminology and wording, in order to gloss over details that are not relevant here, or to use language more compatible with my own usage. I do not think that any such changes affect the substance of Wallace's proof, but will offer apologies in advance, in case that they do.[29]

3.3.10.2 Definitions

Definition 3.18. An "agent" is a decision-making entity who perform "actions" as a result of "decisions" that the agent makes, where each decision chooses amongst a set of "available acts", or at least assigns a preference ordering on those acts.

Definition 3.19. A "microstate" is a quantum state corresponding to a subspace (in some decomposition of identity on the Hilbert space) of our system, where the fineness of grain (the resolution) of the decomposition is the finest that is possible (or at least relevant) to us in the current context. Note that unlike in classical contexts, no claim is made here that microstates are equally probable. (Note also that this is my reading of Wallace's microstates; Wallace himself does not seem to explicitly define the term.)

Definition 3.20. A "macrostate" is a quantum state corresponding to a subspace (in some decomposition of identity on the Hilbert space) of our system, where the choice of such decomposition is

[29] Thanks go to Jacques Mallah for his deep insights into the Deutsch-Wallace proof, which greatly influenced what follows.

"largely fixed by decoherence", but its fineness of grain, according to Wallace, is underspecified.[30]

Definition 3.21. A "state" (macrostate or microstate) will be said to be "available" if there is an act available to the agent which would leave at least one version of him in that macrostate (or microstate).

Definition 3.22. An "event" is any disjunction of any number of macrostates. This allows one to create an "event space" (the set of all such disjunctions) that is coarser-grained than the macrostate space (the subspace of the Hilbert space).

We need to be able to create a coarser-grained event space for the simple reason that there will always be macrostates that contain features that are irrelevant to the agent's interest (in this case, his interest is in getting rewards). We can view the macrostate space as simply the finest-grained event space consistent with decoherence (and hence, for us, synthetic unity).

Definition 3.23. An event space for which each event contains only macrostates that share the same reward will be called a "reward subspace".

Note that using a straightforward counting measure of probability, based on macrostates and microstates, would mean dividing the number of microstates in our macrostate of interest by the total number of microstates. This would yield:

$$p(E) = \frac{|\{|s_k\rangle \in \Omega : |s_k\rangle \in E\}|}{|\Omega|} \tag{3.93}$$

where E is the event of interest (perhaps the set of available macrostates that result in a certain reward) and Ω is the total number of available microstates. The $|X|$ operator simply returns the size of set X.

But this would be the classical approach to probability (explained in more detail in the next chapter), which Bayesians believe to be (along with frequentism) inadequate. So it is not the approach to quantum probability that Wallace is pursuing. The decision-theoretic approach would have us believe that one must approach things from a more subjective stance. The classical approach, on the other hand, assumes that the microstates are somehow objectively existing countables of the system (this is much closer to my own approach, which will also be developed in more detail in the next chapter).

[30] From the perspective of ASU—but not necessarily Wallace's—we can add that the macrostates must respect synthetic unity, which imposes further constraints on the fineness of grain, since resolving the macrostates to a degree of resolution that assigns the same conscious state to more than one macrostate is clearly too fine, while a resolution that assigns multiple conscious states to one macrostate is clearly too coarse. Any macrostates that are said to result from an act must each contain a coherent continuation of the personal identity of the agent who so acted. Any two microstates, therefore, that contain the *same* version of the agent—with the same conscious state—will automatically be placed in the same macrostate.

An interesting question, to which we will return later, is whether Wallace would have been better off just stopping here—now that he has defined his event space, with macrostates and microstates—and using the classical approach? (If so, there would, of course, be little point to a decision-theoretic Born rule proof.)

3.3.10.3 Axioms

Wallace's axioms fall into three categories:

1. *Richness axioms:* These ensure that the there are adequate degrees of freedom in our system—that our system is "rich" enough—for the proof to go through. In particular, these axioms constrain which acts are available or unavailable to an agent. (These axioms are benign if they simple prevent us from having to worry about whether the proof will work in artificially restricted or trivial cases; but we need to ensure that richness axioms do not go further than this, in effect rigging the proof in favour of a certain result.)
2. *Rationality axioms:* These apply to rational reasoning about probabilities in general (or to any other coherent, quantitative means of preferencing one's actions).
3. *Everettian assumptions:* These make assumptions about rational reasoning, for any agent who already believes himself to live in an Everettian universe, and knows all the state information relevant to making his decisions.

Richness axioms The richness axioms can be divided into two types. The first two axioms address *general* features of action in an Everettian context—these are axioms that we would generally assent to for any Everettian scenario that was not arbitrarily limited. The second two axioms are more contrived, as they set up certain abilities for our agent that are *specific* to making Wallace's proof go through.

General richness axioms

1. *Branching availability*:
   Given any set of positive real numbers $p_1, \cdots, p_n$ summing to unity, an agent can always choose some act which results in n different macrostates, and gives Born measure p_i to the ith outcome.
2. *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 H.

Specific richness axioms

3. *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.

4. *Erasure*:
   Given a pair of non-zero states, ψ and φ, in different events ($\psi \in E$, $\varphi \in F$) in some event space, but in the same reward ($\psi \in R$, $\varphi \in R$) in reward space, there is an act $\hat{U}$ available at E and an act $\hat{V}$ available at F such that $\hat{U}\psi = \hat{V}\varphi$.

Branching availability ensures that there is a sufficient variety of acts available to an agent, such that there will always be an act corresponding to any given distribution of Born probabilities. This will not always be the case in reality, but for the purposes of our proof, it seems benign to assume it. One might imagine that, as the agent, if you were given some arbitrary distribution of Born probabilities to duplicate, that you have some kind of quantum experiment kit that you can whip out and set up an experiment that duplicates those Born probabilities.[31] One can easily imagine such a kit, so there is no reason not to allow our agent to have one at his disposal.

Problem continuity simply assures that our agent's actions are in accordance with the unitary evolution of the wavefunction. One might think that this is unnecessary—of course they are unitary, since we are assuming an Everettian context, and hence unitary evolution. However, this axiom is still necessary, since an agent's *available* actions cannot simply be assumed to be due to the usual unitary evolution of the wavefunction—it is not possible for an agent to view his actions this way, and still have a coherent strategy. If he did, he would be limited to strategies in which he simply decides to choose whatever the laws of nature dictates that he *will* choose (whether that is one specific available act only, or all of them in superposition). Availability has a hypothetical aspect. In order to have a coherent decision strategy, an agent must be able to consider—at least for the purposes of having a strategy—that he has free will (at least in some subjective sense) and is capable of actually choosing *whatever* acts he wishes, amongst those that are available to him, not just those that unitary evolution dictates are possible. Hence, "availability" is not simply "whatever results from unitary evolution". One might imagine trying to build a case against Wallace on the basis of this axiom, but I think it is actually thoroughly benign—to deny that it is benign would require insisting that an agent in an Everettian universe is incapable of having strategies. The very idea of an "agent" must allow the agent to have a useful notion that he is capable of making decisions.

We see also here why we cannot assume that an agent knows the system state in *full* detail, since this would require knowing *himself* in complete detail, which is presumably impossible (given Gödelian incompleteness [93, 176]) and would eliminate the need for any decision-making other than the trivial "do whatever nature says you must". Given, then, that we must assume that more acts are available to an agent than are actually mandated by the unitary evolution of the wavefunction, we need to assume problem continuity, so that the hypothetical actions under consideration do not

[31] Wallace does not call them "probabilities", but rather "quantum weights"; but they are simply the Born probabilities, or weights, or measures, so I will call them Born probabilities—just remember that we are not assuming that these are the *actual* probabilities experienced by an observer, which may be determined by some non-Born rule.

have characteristics inconsistent with the postulates of quantum mechanics (so that they may be considered valid *hypothetical* actions).

Wallace's proof is based on an agent's preferring acts that produce the most reward. Hence, *reward availability* is a necessary and benign assumption—there is no reason we should have to consider situations in which the agent is not capable of performing the acts needed to produce reward (this does not mean all acts will produce reward, just that there are always acts available to the agent that produce the rewards that are under consideration in the proof).

The power of *erasure* gives our agent the ability at any point to erase information in the quantum state in order to make the result of two different hypothetical acts look the same (*i.e.* to yield the same macrostate). Of the richness axioms, this is perhaps the one that most needs defending, although I believe it is benign. It does assume a power for our agent that is rather unrealistic, and some may question it on this basis, especially since the power of erasure, in general, seems to imply the power to create branch mergers (recoherence), which are, in general, gross violations of thermodynamics. To convince ourselves that erasure is benign, we must convince ourselves that it does not create recoherence, at least not in the context in which Wallace invokes it in his proof.

Imagine that we have a spin observation that results in two macrostates, $|up\rangle$ and $|down\rangle$, equally weighted by the Born measure, and about to be measured by an apparatus in state $|a_{ready}\rangle$.

$$\frac{1}{\sqrt{2}}\big(|up\rangle + |down\rangle\big) \otimes |a_{ready}\rangle \tag{3.94}$$

Only, here our "apparatus" will essentially be the reward (or its absence), since we will never look at any gauges or pointer positions. A measurement produces a new state that is (possibly) a superposition of macrostates. Assume that *up* results in reward for our agent, and *down* results in no reward:

$$\frac{1}{\sqrt{2}}\big(|up\rangle + |down\rangle\big) \otimes |a_{ready}\rangle \Longrightarrow \frac{1}{\sqrt{2}}\big(|up\rangle \otimes |a_{reward}\rangle + |down\rangle \otimes |a_{no-reward}\rangle\big) \tag{3.95}$$

Bringing in the environment we have

$$\frac{1}{\sqrt{2}}\big(|up\rangle+|down\rangle\big)\otimes|a_{ready}\rangle\otimes|e_{ready}\rangle \Longrightarrow \frac{1}{\sqrt{2}}\big(|up\rangle\otimes|a_{reward}\rangle\otimes|e_{up}\rangle+|down\rangle\otimes|a_{no-reward}\rangle\otimes|e_{down}\rangle\big) \tag{3.96}$$

where $|e_{ready}\rangle$ is the ready state for the environment when uncoupled from the particle, and $|e_{up}\rangle$ and $|e_{down}\rangle$ are the environmental states correlated with the $|up\rangle$ and $|down\rangle$ observable states, post-measurement.

Recall that our macrostates are defined by decoherence (and, for us, synthetic unity). So the right-hand side above only represents two macrostates (and hence a branching) if the reduced density

matrix

$$\hat{\rho} = \frac{1}{2}\big(|up\rangle\langle up|\langle e_{up}|e_{up}\rangle + |down\rangle\langle down|\langle e_{down}|e_{down}\rangle + |up\rangle\langle down|\langle e_{down}|e_{up}\rangle + |down\rangle\langle up|\langle e_{up}|e_{down}\rangle\big) \quad (3.97)$$

is approximately diagonal, so that

$$\langle e_{down}|e_{up}\rangle \approx \langle e_{up}|e_{down}\rangle \approx 0 \quad (3.98)$$

effectively encoding the orthogonality of the basis states into the environmental states.[32]

Erasure, as Wallace uses it, requires that we erase all evidence of the *up*/*down* result, *except* for its reward. This means we now have the decoherence version of erasure:

$$\frac{1}{\sqrt{2}}\big(|up\rangle \otimes |a_{reward}\rangle \otimes |e_{up}\rangle + |down\rangle \otimes |a_{no-reward}\rangle \otimes |e_{down}\rangle\big) \Longrightarrow \frac{1}{\sqrt{2}}|erased\rangle \otimes \big(|a_{reward}\rangle \otimes |e_{up}\rangle + |a_{no-reward}\rangle \otimes |e_{down}\rangle\big) \quad (3.99)$$

Is the right-hand side here an example of recoherence? Clearly not with those environmental variables in there, since the environment still encodes for which spin was obtained. So even though the observer has "erased" the result of the measurement, to the extent that his limited abilities are incapable of recovering the information, the information is still encoded in his environment, so no recoherence has occurred.

The sticking point here is that one could argue that this is simply *not* erasure, since the result of the measurement was not, in fact, literally erased. The whole point of the above right-hand side for Wallace's proof will be that it is identical to the state that obtains when it is *down* instead of *up* that generates reward. However, this is not literally the case, just approximately so. In an earlier version of the proof [225], Wallace essentially argued that this situation *could* be viewed as erasure—or at least effective erasure—since an agent does not care about effects in the environment that are too small to have any effect on his reward. Hence, Wallace argues that this kind of approximate environmentally-entangled "pseudo-erasure" is effectively equivalent to true erasure for the purposes of the proof.

[32] The reduced density matrix is used here merely as a convenient way of talking about decoherence, as it is the standard mathematical tool for doing so. Wallace's proof does not employ such a device, and I do not mean to imply that it is required here. Recall from Ch. 2 that the standard reduced density matrix formulation of decoherence assumes the Born rule, but that decoherence can be formulated without it.

Granting Wallace this point, at least for the time being, what if we *did* achieve true erasure? Then we could ignore the environmental variables, and we would have the non-decoherence version of erasure:

$$\frac{1}{\sqrt{2}}\,|erased\rangle \otimes \big(|a_{reward}\rangle + |a_{no-reward}\rangle\big) \tag{3.100}$$

This global state is now *precisely* the same, post-erasure, as it would have been if we had associated *down* with reward, instead of *up*, which is what Wallace is looking for. But does this constitute recoherence? There is no recurrence of the global state here, since there is a superposition of reward and no-reward after erasure that did not exist before the measurement. However, recoherence via erasure does not require recurrence (*i.e.,* that the global state be literally returned to the same state as before). It only requires that the trajectories of the independent macrostates before erasure, merge into a single macrostate after erasure. We cannot tell from the post-erasure state, as expressed above, whether this is the case, but we can argue that this state cannot reasonably be said to represent a recoherence. Without the environment, the pre-measurement state was

$$\frac{1}{\sqrt{2}}\big(|up\rangle + |down\rangle\big) \otimes |a_{ready}\rangle \tag{3.101}$$

which represents coherence of the *up*/*down* system, and hence a single macrostate. But the above post-measurement, post-erasure state $\frac{1}{\sqrt{2}}\,|erased\rangle \otimes \big(|a_{reward}\rangle + |a_{no-reward}\rangle\big)$ involves a superposition of reward/no-reward, which cannot feasibly be interpreted as a single macrostate, since to do so would require that we posit an observer *who erases his own memory of whether he received reward.*

However, it is not realistic to imagine a human brain that has split into two observers like this without decoherence, so arguing that recoherence has not occurred, based on the observer's memory states, is tantamount to implicitly invoking environmental decoherence, and we are back to not having literal erasure.

However, Wallace may still be fine here, so long as we are okay with his claim that the observer should not care about the difference between approximate erasure that does not erase reward, and literal erasure. We might be tempted to grant Wallace this assumption, if we could not think of any counter-examples. However, we can see from the decoherence and non-decoherence versions of erasure above, that different erasures that Wallace counts as effectively equivalent, may contain different numbers of macrostates (or branches or worlds or observers). The decoherence erasure shown above assumed that there was a single environmental variable, with a single value that is correlated with *up* and with a single observer/macrostate. However, depending on how the erasure is accomplished, there may be many different such variables with a large number of different states that are correlated with *up* and with *different* observers/macrostates. All these situations will produce identical erasures with respect to reward, and according to Wallace the observer is indifferent to which

kind of erasure is accomplished. The exact erasure is just one (theoretical if impractical) example of a huge number of different possibilities, all of which produce a *different* branching structure, with different numbers of observers (or worlds).

Thus, erasure here is benign, but *only* if the observer is indifferent to branching, and the problem here is that branch-counting is exactly what both Everett and the Born-rule objectors *assume* as the obvious *a priori* probability rule. However, branching indifference is itself *another* one of Wallace's assumptions, independent of erasure, which we will examine in its turn below. Given that fact, we can allow Wallace erasure here, but only because branching indifference is assumed elsewhere.

Generic rationality axioms

1. *Ordering*:
   Each act available to an agent can be assigned a "total" ordering, meaning essentially that we can consider that there is a measure on the acts that can be ordered according to the familiar $<$, $>$, $\leq$, $\geq$ and $=$ relations.

2. *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_i s to performing U followed by the V_i's.

The ordering assumption is benign, as it is straightforwardly necessary if we assume that the agent has *some* coherent, quantitative means of reliably evaluating his preferences for different acts.

Diachronic consistency also seems straightforward and benign, as it simply protects us from having to consider pathological cases where the agent arbitrarily changes his preferences after an act has been performed.

Everettian rationality assumptions

1. *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 who prefers U to V given that the initial state is ψ should also prefer U' to V' given that the initial state is ψ'.

2. *Solution continuity*:
   If at some state ψ, act $\hat{U}$ is preferred over act $\hat{U}'$, then sufficiently small permutations of $\hat{U}$ and $\hat{U}'$ will not change this.

3. *Macrostate indifference*:
   An agent doesn't care what the microstate is, provided it's within a particular macrostate.

4. *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 simply prevents us from preferring anything that is not ultimately based on the physical state. To deny this would clearly either be irrational, or deny wavefunction realism. I will therefore consider state supervenience to be benign in an Everettian context (it might not be benign to a pure subjectivist, but Wallace is not taking a pure subjectivist stance).

Solution continuity seems also to be benign. This prevents us from claiming that an agent makes preferences based on arbitrarily small, infinitesimal, distinctions (for instance, without this assumption, someone might try to claim that the boundary lines between categories that mattered to the agent have a fractal structure). If we assume, with Everett, that we can treat agents as discrete automata containing finite information, then solution continuity follows, since the preferencing of infinitesimal distinctions would require an infinite amount of information to be processed in the agent's mind.

The final two Everettian rationality assumptions are the "indifference" assumptions. Macrostate indifference is straightforward, since the macrostates are defined according to decoherence principles, so that the worlds/branches/observers are, in fact, defined by the macrostates, as emergent properties. Thus, it makes no sense for an observer to distinguish between which microstates he may be in, within a macrostate. In fact, while decoherence may define the macrostates (albeit only coarsely), how the macrostates are to be resolved into microstates is thoroughly ambiguous. We could, in fact, if we wanted, define the macrostate as containing only one microstate (itself).[33]

However, even if we were to arbitrarily define our microstates, Wallace could still claim that the agent does not care about which microstate he is in, since each one is indistinguishable from the other, from his point of view (recall here, that he *also* does not care about which *macrostate* he is in, if the macrostates are in the same reward—this is not about "objective" indifference (or caring) of the classical kind; this is subjective indifference, and is ultimately defined with respect to rewards).

So we will accept macrostate indifference, but take note: while the observer does not care which microstate he is in, this does *not* mean he does not care *how many* microstates there are in his macrostate (if this were what Wallace is claiming, then the axiom would *not* be benign, since it would preclude any classical treatment of probabilities, and while Wallace is avoiding the classical framework, there is no indication he is trying to outright subvert it).

[33] To put this into ASU terms, the macrostates are defined in terms of decoherence and synthetic unity. This means that, in spite of the problem of the preferred basis, we are assuming that there are (reasonably) well-defined macrostates, resolved at a coarseness determined by synthetic unity, which forces the preferred basis on us. However, if we wish to further refine the macrostates into microstates, there is no longer any decoherence or synthetic unity principles to help us out, so we have exactly the problem of a preferred basis again, if we wish to talk about well-defined microstates. Decoherence only helps us preference a certain basis to the level of coarseness required by synthetic unity. Any further refinement at finer scales hits the same preferred basis problem.

I am far less confident that branching indifference is benign, for reasons we already examined in the section on erasure; in fact, I believe that the whole question of the validity of Wallace's proof largely comes down to whether one accepts this one assumption. On the one hand, it might seem as straightforward as macrostate indifference, with a similar justification. Why should an agent care which microstate he is in, if it does not affect his macrostate? However, being in a different microstate does not affect the branching structure, by definition. As we already saw in the section on erasure, different methods of erasure may produce wildly different branching structures, and hence different observer counts. So branching indifference must be defended differently than macrostate indifference. It is especially important that Wallace defend this axiom comprehensively, since it is already denied in a wide cross-section of the literature, and at least implicitly by most of the Born rule objectors themselves.

Wallace has two justifications of branching indifference [227],

1. physical unrealizability, and
2. incoherence of the branch count.

The first, he ultimately defends in terms of the second, so in reality, it all comes down to his claim that there is no coherent, discernible branch count, even in principle. Without this incoherence argument, physical unrealizability on its own does not suffice. The idea of physical unrealizability is that one could never actually count the number of branches, since branching is ubiquitous, and happening all the time, everywhere, with great prolificacy. However, the whole requirement of physical realizability, in my opinion, is inconsistent with the idea of objective probabilities. The most rational strategy for an agent with complete state knowledge might well be physically impossible to realize in practice, but could still correspond (being the theoretically optimal strategy) to the actual objective probabilities. There is no reason to presume that the objective probabilities must be actually computable by the agent. Physical realizability hence is only a reasonable requirement if we are assuming there are *no* objective probabilities, and Wallace denies that he is doing this (and his state supervenience axiom would seem to preclude it, anyway). And even if he were doing that, then he would have to defend this clearly controversial stance on probabilities, as yet another axiom.

Wallace attempts to respond to this argument as follows:

> "Firstly, we already know there's at least one possible [physically realizable] rational strategy: the Born rule."

But this does not seem right: just because the Born rule is an existing realizable rational strategy, does not mean it is the only one, nor the optimal one, if it does not correspond to the objective probabilities. I would take Lewis's "Principal principle" [128] as a given here: subjective probabilities

must correspond to objective probabilities when the latter are known by the agent. Wallace's proof is entirely based on the claim that he is able to show that there is *only one* rational strategy. If he cannot do this, then the existence of the Born rule as a physically realizable rational strategy does not say anything about whether a *better* strategy might exist that is also rational, but *not* physically realizable. Just showing that the Born rule is an existing realizable strategy does not prove that this other strategy does not exist, and is not better, unless we assume the Born rule is the *only* rational strategy—but that is just what Wallace is trying to prove.

Wallace further argues

> "Secondly, what would it even be for a strategy to be rational, but physically impossible?"

Again, this argument does not work. A rational but physically impossible strategy might be one that requires infinite computation, for instance. There are well-defined mathematical propositions [93, 176] that have definite, objective truth values, yet these truth values would take infinite time to calculate. If the statement of an outcome's probability is of this nature, then the most rational strategy is physically impossible to realize... unless Wallace means to say that, by definition, it is irrational to seek to employ a strategy that is physically impossible. However, this would be a confused standard of rationality to use for his proof, since it would *a priori* reject a strategy that corresponds to the objective probabilities, on the sole grounds that it is unrealizable by an actual physical agent. Yet, if this is his standard of rational action, then it is a standard that is straightforwardly inappropriate as a means for proving a probability rule, although it might be a reasonable standard to apply to a *practical* analysis of decision-making.

Given the dubious nature of the physical unrealizability argument, Wallace would do better to go straight to the incoherence argument. In fact, he does hedge on the possibility that the idea of an unrealizable rational strategy might be made to work. However, he claims it would not matter, since ultimately, the branch count is not merely physically impossible for a actual agent to calculate; the branch count does not even exist, not even in principle, since "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." [227] There simply is no truth to the question of what the "branch count" is, in a given circumstance, according to Wallace:

> This structure has a significant degree of arbitrariness associated with it, primarily in terms of the coarseness of the grain of the macrostates... 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. [227]

While any branch count would clearly be at least extremely complex to calculate—perhaps not possible in practice—I do not buy Wallace's argument that the whole notion of there being any objectively existing branch count, whether knowable or not, is actually incoherent.

My first problem with his argument above is his invocation of continuity. This violates my principle of the discrete default, explained earlier in §3.1.3 in my response to Stapp: if an argument invokes the continuum, we insist that it be reformulated in terms of discrete limits before buying into it. While Wallace may not accept my principle (since it admittedly entails an anti-infinitistic bias), he certainly cannot claim to have adequately defended his axiom, for the purposes of his proof, if he cannot convince a moderate finitist, who *does* follow this principle, to accept his axiom. And unlike with Stapp's argument, I do not see any way of reformulating branching indifference in discrete terms. The mere fact that branching is continuous in the physical model, given the mathematics of decoherence, does not mean that there is no well-defined branch count. Wallace himself admits that, for a certain resolution of the Hilbert space, there *can* be such a count (his definition of macrostates assumes as much):

$$W_n^O(|\psi\rangle) = \text{number of branches for observer O, for a resolution of } n \text{ dimensions} \tag{3.102}$$

So to prove his incoherence argument, Wallace needs to do more than simply point out that branching is continuous. He needs to show that the limit of the discrete branch count, approaching infinite resolution, is infinite:

$$\lim_{n\to\infty} W_n^O(|\psi\rangle) = \infty \tag{3.103}$$

However, this does not follow simply from the continuous nature of branching (assuming we accept Wallace's claim that branching is continuous, which, as we will see shortly, I do not). Keep in mind a primary justification for the principle of the discrete default: continuous quantities in calculus (the primary intuitive drive behind the idea of an objectively real continuum) are formally defined in terms of limits. So when one makes a statement like "such-and-so is continuous," it is difficult to credit that one is not, more formally, making a statement about limits on discrete processes. Wallace might be right that branching is continuous, but this does not mean that the limit of the branch count is infinite or not well-defined.

In fact, I will further argue that the branch count may, in fact, be well-defined (although I am not claiming to prove that it is). I argued in the section on the preferred basis, that contrary to what some believe, I do not believe that the mere analytic machinery of decoherence is enough, on its own, to pick out the preferred basis, without employing synthetic criteria. We need, in addition, synthetic unity. Decoherence explains why some (but not necessarily all) of an observer's environment appears not to have obvious interference characteristics. But it is, more fundamentally, synthetic unity that

picks out a precise preferred basis for all observables. Since (I presume, and Wallace cannot prove otherwise) that conscious mental states can be finitely described as discrete structures (later, we will use algorithms to describe them), and they must therefore contain only a finite amount of information (even if the physics is continuous), it follows that there will be a *specific* coarse-graining of the Hilbert space that corresponds to macrostates that each contain one, and only one, unique conscious version of the observer. Any finer grain, and a single conscious state will cover more than one macrostate; any coarser, and a single macrostate will contain multiple different versions of the observer.

So Wallace is wrong that the coarseness of grain of the macrostates is ambiguous, and he is even wrong that the branching is continuous. It cannot be continuous, if conscious states contain finite information, since the branching is defined by the macrostates, and a finite amount of information in the observer's conscious state means that there can only be a discrete number of macrostates, and a discrete branching structure. It is nonsensical to say that branching is continuous, if minds are discrete, since branching, as Wallace himself repeatedly emphasizes, *is not even a physical process*. It is, rather, entirely perspectival (*i.e.*, synthetic). Hence, so long as we assume that a person's perceptual perspective is defined discretely, it is simply not possible for branching to be continuous.

Wallace makes the mistake, I believe, of pinning too much of the branching structure on the mathematics (the analytic structure) of decoherence alone, and assuming that since decoherence does not precisely define the branching, that it is literally, objectively ambiguous. However, the use of synthetic unity promises to disambiguate the coarseness of grain of the branching, and yield the branch count as a definite quantity (and it does so, recall, for the exact same reasons that it is needed, alongside decoherence, to yield a precise preferred basis, something decoherence also cannot do entirely on its own).

Since so much of the literature assumes branch-counting as an *a priori*, I believe the lack of a convincing argument for branching indifference is ultimately fatal to Wallace's proof. This makes the formal part of his proof, while not pointless, certainly not the most crucial part of his argument. The crux of his argument is rather his informal defence of the branching indifference axiom. And, while Wallace may argue against my acceptance of a discretely defined synthetic unity of consciousness, he certainly has not adequately shown it to be irrational, and so, as a Born rule *proof*, his demonstration fails.

While I have not accepted branching indifference, Wallace's proof is still an important contribution to the debate, since it illustrates why the question of branching indifference is so crucial. In other words, it demonstrates why we must accept that *either* the Born is true *or* the branch-count matters. My own views will be explained in later chapters, but for now, I will simply point out that

I do, in fact, agree with Wallace that branch counting is irrational, just not for the reasons he has given. My own reasons will have to wait for the next chapter, as they depend on my opinions on the interpretation of probability theory.

In the meantime, let us assume, for the sake of argument, that branching indifference holds, and for the sake of completeness, we will look briefly at the structure of the rest of Wallace's proof (which I believe is relatively straightforward).

3.3.10.4 The theorem

Wallace proves three lemmas, followed by his Born rule theorem.

1. *Equivalence lemma:*
   Depends on:
   erasure, branching indifference, ordering (transitivity), diachronic consistency, state supervenience

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

   (*Note:* since this lemma will turn out to be the crux of the whole proof, I will refer to the axioms on which this lemma depends as the "equivalence axioms".)

2. *Nullity lemma:*
   Depends on: equivalence lemma, diachronic consistency.

   An agent is indifferent to a possible outcome of an act if, and only if, that act has zero Born weight.

3. *Dominance lemma:*
   Depends on: equivalence lemma, diachronic consistency.

   Suppose that two acts each only have two possible rewards r_1, r_2 as outcomes, with $r_1 \succ r_2$ 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.

4. *Born-rule theorem:*
   Depends on: equivalence, nullity and dominance lemmas.

   Consider two utility functions: *(1)* one ("reward utility") is a function of the rewards, and *(2)* the other ("Born utility") is a function of the Born weights of the rewards. It then follows that any given act is preferred over any other if, and only if, its expected utility is higher according to *both* of these utility functions. (This allows for only one unique reward utility function, so long as we allow for scaling with an offset.)

The nullity and dominance lemmas are straightforward. Translating them into probabilistic terminology: once one accepts that equal amplitudes mean equal probabilities, it follows straightforwardly that greater amplitudes yield greater probabilities, and that zero amplitude yields zero probability

(and vice-versa). This takes some work to prove, but is a fairly intuitive and believable result, and I believe Wallace's derivation of it is straightforward, so I will forego the details. The Born rule follows straightforwardly from these three lemmas—and this is also intuitive, given that the three lemmas together basically require our probability measure to have the same ordering as rational preferences, which is tantamount to representing the Born rule in decision-theoretic terms.

3.3.10.5 Erasure and equivalence

The equivalence lemma is the real crux of the whole proof. Wallace attempts to prove equivalence by showing that it is simply not rational to tie rational preference (and hence, ultimately, probability) to a feature of the wavefunction other than amplitude, as one can then imagine situations in which *identical* wavefunction states resulting from two different strategies, could be assigned *different* preferences. So long as the Born ordering is violated, then for any two potential global system states with *different* preferences, it will always be possible to imagine erasing information from one of the states to make the two states *identical* (as already discussed, it may not really *always* be possible to actually do this, but it certainly seems that we can set up situations where it is possible, and that should be enough to make the argument stick). If I know that doing A and doing B will result in the exact same global state, there can be no justification, given state supervenience, for preferring one of A or B over the other. This clever trick, using erasure, is really at the root of Wallace's formal proof. Unfortunately, as we have seen, the process of erasure can produce any number of different branch counts, and so the agent cannot really be indifferent to it unless we eliminate branch counting as significant. As already discussed, Wallace simply assumes as an axiom that branch counting is irrational, and provides informal, verbal arguments, which I have already rejected.

Nonetheless, allowing branching indifference for the sake of argument, let's move on. Wallace demonstrates equivalence by simply *erasing* the distinguishing information (the non-mutual information) encoded in the two global states, to make them indistinguishable. The trick, of course, is that he has to do this in such a way that the two hypothetical global states start out with clearly different preferences, *and* so that the erasure process itself is guaranteed not to have any influence on preference. If he can do this, he will have shown that identical states *must* have identical preferences, from which he can argue that, more generally, identical amplitudes must have identical probabilities.

To convince us of this, Wallace's primary tool is his thought experiment of two games (game 1 and game 2) involving betting and erasure. In game 1, you are presented with a quantum state (the details are known to you) prepared in a superposition of "up" and "down", with equal amplitude for each. When you make an observation, the entire (global) system will be in a superposition of your seeing "up" and your seeing "down". A measuring apparatus looks at the result, and if it sees "up",

then you get a payout, if "down", you receive nothing. Game 2 is exactly the same, except that "down" results in payout, and "up" does not. Your own observation of the result is only through getting the payout or not, not direct observation of the result as "up" or "down", which is never directly revealed to you. Both terms in the superposition have amplitude $\frac{1}{\sqrt{2}}$ (and hence Born measure of 50%). You will be asked to place a bet on the outcome, and the question at hand is simply this: which game should you prefer, 1 or 2? It may seem self-evident that there should be no preference, since the games are identical to each other, except for the different "up" and "down" spins, which have equal amplitude. However, this would be to assume amplitude dependence. (Perhaps there is some built-in preference for "up" spins, with respect to observation.)

Two games:

$$\textit{Game 1}: |\psi> = \tfrac{1}{\sqrt{2}}(|up;\, reward\rangle + |down;\, no\, reward\rangle)$$
$$\textit{Game 2}: |\psi> = \tfrac{1}{\sqrt{2}}(|up;\, no\, reward\rangle + |down;\, reward\rangle)$$

Now imagine a second version of each game, where the up/down state is erased, immediately after you receive the payout. Your payout is not erased, but all record of whether it was "up" or "down" that resulted in your payout (or lack of it) is erased. Post-erasure, then, we have the following situation:

Erasure:

$$\textit{Game 1}: |\psi'> = \tfrac{1}{\sqrt{2}}(|erased;\, reward> + |erased;\, no\, reward>)$$
$$\textit{Game 2}: |\psi'> = \tfrac{1}{\sqrt{2}}(|erased;\, no\, reward> + |erased;\, reward>)$$

Post-erasure, the two states of the systems are thus indistinguishable. Therefore, Wallace argues, according to state supervenience, you *must assign them the same preference,* and the principle of equivalence is proved.

Of course, in this example, the Born rule and branch counting yield the same results. Let's look at an example where they do not. Imagine that, in certain situations, we perform a second *up/down* observation, and we set up the following:

Now we have

$$\textit{Game 1}: |\psi> = \tfrac{1}{\sqrt{4}}\,|up, down;\, reward\rangle + \tfrac{1}{\sqrt{4}}\,|up, up; reward\rangle + \tfrac{1}{\sqrt{2}}\,|down; no\, reward\rangle$$
$$\textit{Game 2}: |\psi> = \tfrac{1}{\sqrt{4}}\,|down, up; no\, reward\rangle + \tfrac{1}{\sqrt{4}}\,|down, down;\, no\, reward\rangle + \tfrac{1}{\sqrt{2}}\,|up;\, reward\rangle$$

By observer-counting, we should prefer game 1 (two-thirds versus one-third chance of reward). By the Born measure, we should be indifferent (the chance is one half either way).

After erasure:

$$\textit{Game 1}: |\psi> = \tfrac{1}{\sqrt{4}}\,|\textit{erased}; \textit{reward}\rangle + \tfrac{1}{\sqrt{4}}\,|\textit{erased}; \textit{reward}\rangle + \tfrac{1}{\sqrt{2}}\,|\textit{erased}; \textit{no reward}\rangle$$
$$\textit{Game 2}: |\psi> = \tfrac{1}{\sqrt{4}}\,|\textit{erased}; \textit{no reward}\rangle + \tfrac{1}{\sqrt{4}}\,|\textit{erased}; \textit{no reward}\rangle + \tfrac{1}{\sqrt{2}}\,|\textit{erased}; \textit{reward}\rangle$$

But we can't say anything about number of observers, until we note that this is really

$$\textit{Game 1}: |\psi> = \tfrac{1}{\sqrt{2}}\,|\textit{erased}; \textit{reward}\rangle + \tfrac{1}{\sqrt{2}}\,|\textit{erased}; \textit{no reward}\rangle$$
$$\textit{Game 2}: |\psi> = \tfrac{1}{\sqrt{2}}\,|\textit{erased}; \textit{no reward}\rangle + \tfrac{1}{\sqrt{2}}\,|\textit{erased}; \textit{reward}\rangle$$

Now the two games, once again, lead to identical results, and the number of observers that get a reward is now one half (yielding the same result as the Born measure). Hence, the Born measure is resistant to erasure, whereas observer-counting is not. Clearly it seems irrational that erasing something that has nothing to do with my payout, after the fact, should change my strategy. Or, at least, this is Wallace's argument.

This is considerably cleaner than Wallace's previous attempts [225], in which equivalence was not a theorem, but another assumption in addition to branching indifference. Here, Wallace has isolated the *a priori* "rational" aspect of equivalence into the erasure axiom, which is relatively benign compared to his old "equivalence axiom". This reduces the potentially controversial axioms from two to only one—namely, branching indifference. Of course, if we reject branching indifference (or at least Wallace's defence of it), the whole line of reasoning above falls apart, since we now have a huge number of different possible branch counts for situations that the observer is not supposed to distinguish between, according to Wallace.

Discussion Whether or not we accept that Wallace has genuinely proved the Born rule, it is an illuminating proof that cleanly separates out the formal deductive aspects of the argument from the more verbal and intuitive aspects (his defence of the axioms). It is also concerned with actually directly demonstrating that amplitudes are what matter, rather than proving that they reduce to some other standard that remains itself unjustified.

Compare the proof, for instance, to Everett's or Hartle's proofs, both of which have much stronger amplitude-dependence assumptions, with no systematic attempt at justification. Everett assumes (as Hartle does more implicitly) that the correct probability measure is a function of amplitude, and then assumes *a priori* that the measure must conform to branch counting, without giving a trace of justification. Wallace managed in [225] to weaken amplitude-dependence to merely the assumption that equal amplitudes yield equal probabilities (his equivalence axiom), and then he weakened this still further in [226, 227] to the erasure axiom, which no longer even has the obvious character of an amplitude-dependence assumption.

However, Wallace still must assume branching indifference, which in combination with erasure, yields equivalence and hence amplitude-dependence. So one might argue that amplitude-dependence is still assumed, in a weak sense, since branching indifference indirectly supports amplitude-counting by denying its main competitor, branch-counting (although this does have the advantage of eliminating third options, that count neither branches nor amplitudes, whether one accepts branching indifference or not).

There is another possibility here, and that is the possibility that Wallace's proof simply translates Gleason's proof into a decision-theoretic framework. Wallace addresses this issue somewhat in his response to the charge that contextual measurement provides an alternative rational strategy to the Born rule, even in light of his proof (and recall that Gleason assumes noncontextuality). Wallace responds [227] by pointing out that any contextual probability rule violates state supervenience, and so in his scheme is eliminated for that (clearly rational) reason. A contextual probability rule would have to be one that gave different results depending on how the observer was interpreting the *same* state. Hence, the rule would depend on more than just the physical state of the system, and would thereby violate supervenience, which Wallace takes (I think rightly) to be irrational.

However, Wallace's response here simply strengthens the suspicions that his proof is, in essence, simply a restatement of Gleason, since he admits that his state supervenience implies noncontextuality, which is all that is needed to prove the Born rule via Gleason. But, if this is so, it is unclear why we need the elaborate decision-theoretic framework just to patch up this one aspect of Gleason's proof. Why not, then, just add a state supervenience axiom to Gleason's proof, which very directly yields noncontextuality and we are done—no need to delve into decision theory at all. State supervenience itself is not even a particularly decision-theoretic notion, and could be transplanted in this fashion rather straightforwardly. If one is a hard-core Bayesian, of course, one may have independent arguments—in term of probability theory—for preferring the decision-theoretic approach (and we will save a detailed discussion of these issues for the next chapter). However, for the rest of us—especially those who believe in objective probabilities—the whole framework of decision theory would seem to be unnecessary window-dressing, unless we can show a clear incentive for using it. If "state-supervenience + Gleason = Wallace", as Wallace himself seems to imply, then the Deutsch-Wallace proof does nothing new. In fact, Gleason's proof would seem to be a much stronger result, since it only requires noncontextuality (or state supervenience, if we like) and does not require all the additional axioms that Wallace uses, such as branching indifference.

But perhaps the notion of "rationality of action" *is* helpful here: a purely mathematical approach would never tell us *why* state supervenience is disallowed. It is disallowed in Wallace's proof because it is an irrational strategy for decision making. This does tend to lend Wallace's approach

more credibility, since it motivates state supervenience. However, it is still not clear to me that the decision-theoretic framework is necessary, or even preferred. State supervenience seems to me a very straightforward assumption to make within a non-Bayesian and objective view of probabilities, given the standard Everettian axioms of wavefunction realism, psychophysical parallelism and servomechanism equivalence. If probabilities are objective in such a system—as they are in ASU—state supervenience follows, at least as strongly as it does for Wallace. Except that, in this objectivist framework, it becomes a rational *metaphysical* axiom, rather than an axiom about rational decision-making. For, if quantum probabilities are objective (and exist for bunny rabbits as surely as for humans), and our subjective experience is wholly a result of the mechanistic actions of servomechanisms (feedback control devices), which are explained in entirely physical terms (psychophysical parallelism), then it seems unavoidable that probabilities must depend only on the quantum state of the system (state supervenience)—at the very least, this argument is no weaker than Wallace's own decision-theoretic defence of state supervenience.

So, unless we have philosophical reasons to insist on subjective probabilities (what I am calling "hard-core" Bayesianism), the benefits provided by the decision-theoretic approach are unclear, in light of Gleason's theorem, even if we were to accept branching indifference and a subjectivist version of state supervenience. Of course, Wallace does not present his theorem as useful only for hard-core Bayesians. Many proponents of decision theory (Wallace claims the majority, in fact) are *not* hard-core Bayesians, and, when decision theory allows more than one conflicting rational strategy, many decision theorists are happy to require us to choose the one that is in line with the "objective" probabilities. They do not then necessarily conclude that the two strategies are "equally valid". But Wallace claims that, in fact, his decision theoretic proof does *not* leave us with more than one rational strategy to choose from—it leaves us *only* with the Born rule. Thus, a hard-core Bayesian might still find his approach very appealing: it claims to derive an objective decision-making rule, essentially equivalent to the Born rule, without assuming there is any such thing as objective probabilities. However, since Wallace himself does not in any way try to claim that quantum probabilities are *not* objective—and recall my position in Ch. 1 that it is almost impossible to argue that quantum probabilities are *not* in fact objective—the question remains: why decision theory?

My own approach has been the synthetic *a priori* approach, and synthetic unity. This approach is essentially a version of the anthropic principle, which is controversial in its own right (details to be discussed in later chapters), so it will not convince everyone, but for those who share my assumptions, and find them to be constitutive of rationality, this approach will be preferred over Wallace's. And if we can get to Wallace through Gleason plus one axiom that fits naturally into the framework of ASU, then this is a much more direct route than using a decision theoretic framework

that seems like excessive window-dressing.

This also raises the issue of whether we might be able to use some other combination of Wallace's axioms to weaken the amplitude-dependence of Everett's stage 1 proof, or Hartle's proof, so that they are essentially as strong as the modified-Gleason or Wallace proofs. For instance, since Everett's proof falls down by assuming amplitude-dependence in stage 1, and branch-counting in stage 2, it would seem that by dropping stage 2, and adding Wallace's equivalence axioms (with the decision-theoretic window dressing removed), we should be able to produce a competitive, and simpler proof. This proof would have much more in common with Wallace's existing proof than would the modified Gleason proof—the point here is not that we are throwing out Wallace's proof, but that we are divesting ourselves of its decision-theoretic trappings, which do not seem to be essential, other than to convince people who are already convinced of hardcore Bayesianism. I am not claiming this proof would convince the Born rule objectors, but it might do as well as Wallace's.

There is so much of Wallace's *formal* proof that depends on the *informal* defence of his rationality axioms, and these informal arguments are of such a general nature, and not dependent on the decision-theoretic framework, that it seems to me that the "real work" of the proof is being done by these informal arguments, that lie outside decision theory, and so the question naturally arises whether the part of Wallace's proof that is undeniably an important contribution—the significant weakening (compared to Everett) and rational justification (compared to Gleason) of the required axioms—could not be translated into a non-decision-theoretic framework. Of course, it is not really for us to blame Wallace for not doing this. He is working within a decision-theoretic framework—definitely the preferred framework for at least some people. One has to work within some framework. It is the job of others now to see if an equally-strong proof cannot be constructed outside of decision theory; expecting Wallace to guarantee us that this is (or is not) impossible would be expecting too much. I would still argue, however, that branching indifference needs a better justification than that given by Wallace. However, it does not seem fruitful to spend time looking for such a thing, since it appears that state supervenience implies Gleason, and we can get to the Born rule that way without any direct repudiation of branch-counting.

In spite of my criticism, I actually think Wallace's conclusions are basically correct: branching *itself* cannot matter *per se* (since worlds and observers are not the fundamental ontic entities of the system). However, I do not believe this for quite the same reasons as Wallace, and I am suspicious of the idea that decision theory is necessary, or even very helpful, in making these points, or that the Born is rule is any more "proven" now than it was before, given that we already had Gleason's theorem, which would seem a stronger proof based on much weaker assumptions. It seems that the validity of Wallace's proof hinges less on the adoption of decision theory, and more on whether one

adopts something like Strong AI (or servomechanism equivalence) and synthetic unity (or "subjective uncertainty", the analogous concept in Wallace's system). Or, more generally, it depends on how one defines "observers" and whether one thinks that counting them (or their worlds) is the way to compute probabilities in the first place.

3.3.11 Zurek's Classical Objectivist Proof

Zurek has recently presented a derivation of the Born rule [239, 238, 237, 240, 188] based directly on the properties of environmentally-induced decoherence. Although he is more agnostic about many-worlds than an out-right advocate of the MWI, his work is not only consistent with many-worlds, but he identifies his approach as broadly a "relative state" approach, and so his proof can be reasonably categorized as an Everettian response to the Born rule objection.

Zurek's approach has much in common with the approach I am taking. While it makes its own assumptions, it has the advantage over Wallace's approach (at least from our perspective here) that these assumptions are all situated squarely in the objectivist probability camp: quantum probabilities are objective in nature and can be derived from the physical configuration of the system, based on symmetries inherent to that system—in other words, based on an objectively derived classical principle of indifference.

This is in line with the "generative interpretation" of probability that I will advocate for in Ch. 4: probabilities are (or at least can be) objective features of the generating conditions of an experiment, not merely empirical frequencies of outcomes, or measures of uncertainty, or descriptions of subjective states of belief.

Zurek's proof, while based on the ideas of decoherence, doe not employ the usual density matrix framework for decoherence. Recall from Ch. 2 that this framework already assumes the Born rule, so to use it in a Born rule proof would commit the sin of circularity so common in Born rule proofs. But recall also from Ch. 2 that the use of reduced density matrices in decoherence theory is a convenience, not an absolute necessity. Since the objective situation can still be described in terms of pure states, it should be possible in principle to talk about decoherence without the (admittedly convenient) tool of reduced density matrices. This is more or less what Zurek attempts to do.

Zurek frames his proof in terms of what he calls "Quantum Darwinism", as he uses an evolutionary metaphor to explain it (and even suggests that it may be more than just a metaphor). I am unsure that I buy into this analogy, and so I will not employ it here. Schlosshauer and Fine [188] have critiqued Zurek's proof, and although they find it very promising, they have isolated what they believe are several assumptions of the proof. They have also clarified the exposition of the proof in a number of ways, so I will generally (but not entirely) follow their exposition of the proof in what

follows.

Zurek starts by dividing the universe into a system S represented in Hilbert space H_S and its environment E (which is just the rest of the universe), represented in Hilbert space H_E. A measurement is seen as a communication of information from S to E:

$$|s_k\rangle\,|\mathrm{e}_0\rangle \Longrightarrow |s_k\rangle\,|\mathrm{e}_k\rangle \tag{3.104}$$

where $\{|s_k\rangle\}$ are basis states for S, and $\{|\varepsilon_k\rangle\}$ are basis states for E.

Consider first the general composite state where the subsystems' Hilbert spaces are 2-D, and the amplitudes' norms are equal, but possibly with different phases:

$$|\psi_{SE}\rangle = \frac{1}{\sqrt{2}}\left(e^{i\alpha_1}\,|s_1\rangle\,|e_1\rangle + e^{i\alpha_2}\,|s_2\rangle\,|e_2\rangle\right) \tag{3.105}$$

While this is a fairly restricted example, it is enough to address the substance of a Born rule proof. Recall from Wallace's proof that, once equal probabilities for equal amplitudes are accepted (*i.e.,* we have demonstrated amplitude dependence), proving the Born rule for the more general case follows easily (although we will start with a slightly more general case, with equal amplitude norms but arbitrary phase factors). Likewise, the restriction to 2-D is not substantial, but will be easily generalized to higher dimensions later in the proof. I will therefore focus just on the 2-D equal-amplitude case, for ease of exposition.

Now imagine an arbitrary pair of unitary transformations, each acting effectively on just one of the subsystem Hilbert spaces:

$$\begin{aligned} \hat{U}_S &= \hat{u}_S \otimes \hat{I}_E \\ \hat{U}_E &= \hat{I}_S \otimes \hat{u}_E \end{aligned} \tag{3.106}$$

Definition 3.24. The composite state is said to be "envariant" (from "environmentally-assisted invariance") under $\hat{U}_S$ and $\hat{U}_E$ if it is invariant under the application of both $\hat{U}_S$ and $\hat{U}_E$.

$$\hat{U}_E(\hat{U}_S\,|\psi_{SE}\rangle) = |\psi_{SE}\rangle \tag{3.107}$$

This means that there is an operation on the environment that is capable of "undoing" an operation just performed on the system. Imagine that the operation on the environment is supposed to be an observation or at least measurement-like interaction, such as the registering of information via a change in a pointer state. Envariance would mean that the registering of the pointer state is capable of "undoing" the change in the system. Likewise, repeating the operation on the system would have the affect of undoing the pointer change in the environment. Clearly, this cannot really be a measurement operation. Recall that, in the decoherence picture of measurement, measurements

generally take place in the context of an environment that is inaccessible to the observer, so, in fact, there is no way an observer could ever know what the envariant features of the composite system actually are, and their (local) description of S will be independent of these features.

Zurek uses the example of an operation that only shifts the phase of the Schmidt terms, so we get:

$$\hat{U}_S \left|\psi_{SE}\right\rangle = \frac{1}{\sqrt{2}} \left(e^{i(\alpha_1+\Delta\alpha_1)} \left|s_1\right\rangle \left|e_1\right\rangle + e^{i(\alpha_2+\Delta\alpha_2)} \left|s_2\right\rangle \left|e_2\right\rangle \right) \tag{3.108}$$

with the action of $\hat{U}_E$ capable of undoing this phase shift (simply by performing the negative of the same shift) and leaving us back with the original state:

$$\begin{aligned} \hat{U}_E \left(\hat{U}_S \left|\psi_{SE}\right\rangle \right) &= \frac{1}{\sqrt{2}} \left(e^{i(\alpha_1+\Delta\alpha_1-\Delta\alpha_1)} \left|s_1\right\rangle \left|e_1\right\rangle + e^{i(\alpha_2+\Delta\alpha_2-\Delta\alpha_2)} \left|s_2\right\rangle \left|e_2\right\rangle \right) \\ &= \frac{1}{\sqrt{2}} \left(e^{i\alpha_1} \left|s_1\right\rangle \left|e_1\right\rangle + e^{i\alpha_2} \left|s_2\right\rangle \left|e_2\right\rangle \right) \end{aligned} \tag{3.109}$$

Thus, the composite state is envariant under this transformation pair (and in general, under any phase shift). This means that an observer of S can never know about these phases. They are not local properties of S, but global properties of the entangled composite system. Hence, there is no way that such envariant features can affect the probabilities of the Schmidt states, and the original state will have the same probabilities assigned to its terms, no matter what the phases α_1 and α_2, including when they are zero:

$$\left|\psi_{SE}\right\rangle = \frac{1}{\sqrt{2}} \left(\left|s_1\right\rangle \left|e_1\right\rangle + \left|s_2\right\rangle \left|e_2\right\rangle \right) \tag{3.110}$$

This should be expected, of course, as one would not expect phase differences to result in different probabilities; however, it sets us up to perform the same trick for cases that are less obvious. We will henceforth assume that $\alpha_1 = \alpha_2 = 0$, which brings us to the point where we really do only need to show that equal amplitudes mean equal probabilities in order to prove the Born rule (since we have already shown that the phase factors do not matter).

Zurek next introduces a second kind of envariant transformation: the "swap", which switches around which states of the environment are correlated with which states of the system:

$$\hat{U}_S \left|\psi_{SE}\right\rangle = \frac{1}{\sqrt{2}} \left(\left|s_2\right\rangle \left|e_1\right\rangle + \left|s_1\right\rangle \left|e_2\right\rangle \right) \tag{3.111}$$

with a "counter-swap" on the environment putting things back the way they were:

$$\hat{U}_E \left(\hat{U}_S \left|\psi_{SE}\right\rangle \right) = \frac{1}{\sqrt{2}} \left(\left|s_1\right\rangle \left|e_1\right\rangle + \left|s_2\right\rangle \left|e_2\right\rangle \right) \tag{3.112}$$

As before, with the phase transformations, we assume that an observer of S has no effective access to the environment. Hence, the correlation of an $\left|s_k\right\rangle$ relative state with an $\left|e_l\right\rangle$ relative state is

envariant *for such an observer.* And, of course, this means that there is no way that $|e_l\rangle$ could possibly be a pointer state said to "measure" the state of the system $|s_k\rangle$. How these relative states correlate is a feature of the composite system.

Now, since which relative state correlates with which relative state can be completely swapped around, this swapping must be regarded the way we regarded phase shifting: it can never be thought of by the observer of S as a local property of S, since inaccessible changes in the chaotic, complex environment can "undo" the correlations. Hence, the different terms in the Schmidt decomposition must be equally probable.

Of course, while this argument has some intuitive appeal, it is still not proven, and, in fact, is identified by Schlosshauer and Fine [188] as the major, and most problematic, of Zurek's assumptions:

Assumption 3.25. *Probabilities associated with the Schmidt states $|s_k\rangle$ of a system S entangled with another system E remain unchanged under the application of an envariant transformation $\hat{U}_E = \hat{I}_S \otimes \hat{u}_E$ that only acts on E (and similarly for S and E symmetrically exchanged):*

$$\begin{aligned} p\left(|s_k\rangle\,;\,\hat{U}_E\,|\psi_{SE}\rangle\right) &= p\left(|s_k\rangle\,;\,|\psi_{SE}\rangle\right) \\ p\left(|s_k\rangle\,;\,\hat{U}_S\,|\psi_{SE}\rangle\right) &= p\left(|e_k\rangle\,;\,|\psi_{SE}\rangle\right) \end{aligned} \tag{3.113}$$

If we grant Zurek this assumption then his Born rule proof may follow through (Schlosshauer and Fine identity three other assumptions, as well, but they are relatively benign compared to this one).

The assumption is not without its appeal. In the pre-measurement state, a composite system's superposed subsystem states will display envariance. Zurek sees this envariance as the symmetry that is required in classical probability in order to have a "principle of indifference": a means of determining what entities are equally probable (which entities to "count"). Given equal amplitudes, we have equal probabilities, since the swap makes no difference to the observer, and cannot even be considered by him/her to be an actual property of the measured system, in the first place. Hence, the probabilities (pre-measurement) must be equal.

> "When the state of the observer's memory is not correlated with the system, and the absolute values of the coefficients in the Schmidt decomposition of the entangled state describing SE are all equal, and E cannot be accessed, the resulting state of S is objectively invariant under all local measure-preserving transformations. Thus, with no need for further excuses, probabilities of events $\{|s_k\rangle\}$ must be—prior to measurement—equal. [237, p12]

Schlosshauer and Fine argue, however, that there is a gap in Zurek's defence of this assumption. All that can really be concluded from envariance is that the correlations and their probabilities are inaccessible to a local observer of S. This is far short, they claim, of an objective principle of

indifference for states $|s_k\rangle$, which could still be of different probabilities, even given their indifference from the perspective of the observer of S.

Once measurement is made, of course, then information is imprinted in E about S (they share mutual information). Now, the envariance is destroyed, the symmetry is broken, and the observer may measure the pointer state, even (possibly) with certainty.

Suppose there are states of S (say, $|u\rangle$ and $|v\rangle$) that produce an imprint in E_A, a subsystem of E (which plays a role of an apparatus), but remain unperturbed (so they can produce more imprints). This repeatability Zurek takes as a fundamental requirement for an "objective reality" (although he seems to mean by this something more like "stable environment"). This repeatability implies:

$$\begin{aligned} |u\rangle|e_0\rangle &\implies |u\rangle|e_u\rangle \\ |v\rangle|e_0\rangle &\implies |v\rangle|e_v\rangle \end{aligned} \tag{3.114}$$

where $\{e_k\}$ are basis states for the apparatus. Since unitary evolution preserves the inner product:

$$\langle u|v\rangle = \langle u|v\rangle\langle e_u|e_v\rangle \tag{3.115}$$

and where $\langle e_0|e_0\rangle = 1$.

This means that either

- $\langle e_u|e_v\rangle = 1$ (the transmission of information failed), or
- $\langle u|v\rangle = 0$ (the transmission succeeded, and the states are orthogonal).

Thus, when information is transferred from the system to its environment, in such a way that it can be repeated, with the same result, the states of the system that can produce such imprints must be orthogonal. In addition, this means that the imprint can be repeatedly made in independent fragments (subsystems) of the environment.

The astute reader will notice that this is essentially an argument based on the ideas of environmental decoherence from §2.5.4, but developed here outside of the context of reduced density matrices, where the Born rule is assumed.

Since the environment is not in any practical sense accessible to the observer, it is only by massive redundancy that the observer's measurement can be really said to be *objectively* fixed by the state of the environment. This is the "Darwinian" part of Zurek's scheme, as the more redundant copies are said to "proliferate" more and be "fitter" in the environment. I find the evolutionary metaphor here difficult to take too seriously. I am not sure that I see the requisite notions of reproductive fitness and competition that would be needed to make the analogy work as anything more than a suggestive metaphor.

However, a connection can be made between the "redundancy" of Zurek's imprints, and my algorithmic interpretation. A high degree of compressibility is associated, in algorithmic information theory, with a short description length and a high probability. Hence, "many different imprints" might be taken as another way of saying "high information-theoretic probability", which implies environmental stability (§6.8).

However, the algorithmic interpretation is still quite different from quantum Darwinism, as the algorithmic probability calculations we will perform later in Ch. 6 are all based on synthetic unity, and so are measures of the actual state of the observer's brain, rather than direct measures of redundancy in a physical environment.

But is synthetic unity even required, then, to justify the Born rule, given Zurek's derivation? I think it is, or at least it is just *as* required as it was (in §3.1.3) to justify the choice of a preferred basis. Recall from our discussion of that issue that we concluded that decoherence accounted for the preferred basis, *in a sense*, but only if we interpreted the requirements of the decoherence program as implying a synthetic *a priori* principle something like synthetic unity. Otherwise, the decoherence program could be construed as falling prey to Stapp's objections.

Likewise, with Zurek's Born rule derivation. There are clearly a few assumptions about measurement that Zurek makes, in order to get his proof. One of the most crucial ones is that the observer and system are isolated from the environment subsystem. If we allow for a decomposition that includes an *observer* subsystem, as well as system and environment, these issues become, I think, clearer than they are in Zurek's analysis. Recall that we established that how the envariant features of S were correlated with the environment depends on the overall, entangled system, and cannot be a *local* feature of S. However, how they may or may not be said to be correlated could also depend on some *other* alternative decomposition. In particular, it will depend on the nature of the observer subsystem. It is possible there could be a new decomposition that will take *part* of the previous environment and consider it a new kind of observer, with extreme information processing capacity, for which the envariance symmetry will already be broken, before the previous symmetry-breaking measurement is even made. Hence, it is difficult to see how we can talk about "objective reality" or "environmental stability" evolving out of redundancy of imprinting, without also placing this within the context of synthetic unity.

However, as discussed in §3.1.3, the entire decoherence scheme could be read in such a way so as to ultimately require synthetic unity, by its own lights, and so none of this is really a total rejection of Zurek's solutions to either the preferred basis problem *or* the probability problem. But even in that case, since ASU is clearly, itself, another assumption, Zurek's proof would still not convince everyone. However, Zurek's proof and his whole decoherence framework is highly resonant with the

algorithmic account of quantum probabilities I am trying to develop, and his Born rule proof, while still dependent on some assumptions, may well be (in my opinion) correct in its essentials.

3.3.12 Gleason's Proof

3.3.12.1 Gleason Noncontextuality

Gleason's proof [91], published the same year as Everett's, is another proof of the Born rule, but is quite differently motivated from the Everett-style frequentist proofs. It makes its own unjustified assumption[34], namely that measurements are "non-contextual", so that the probability of a particular outcome is independent of what basis it is considered to be a part of. In other words, we cannot use a norm-squared amplitude measure for one basis, and some other rule for another basis—the probability we get has nothing to do with what kind of measurement we are doing *per se*, or what other measurements we might be simultaneously performing; it depends only on the wavefunction and the outcome state in question. This may seem intuitive, but it is unclear whether it is any more intuitive than amplitude-counting or branch-counting.

Noncontextuality is not a given from the wavefunction alone, and so is generally considered an unjustified assumption. However, there are still many who consider it self-evident, who point out that simply because an assumption is unprovable does not mean it is unjustified. For instance, it is not a given—from the wavefunction alone—that it is irrelevant which outcome Mickey Mouse would consider neater (if only he were real). Yet, for some reason, no one considers this an unjustified assumption (although it clearly *is* an *unproven* assumption).

However, the Mickey Mouse argument clearly does not apply to noncontextuality if one adopts a subjectivist view of probability. Bayesians, recall, consider probability to be primarily a matter of uncertainty or degrees of belief. If this is true, it may be justified to ignore the likely opinions of a possible mouse, but it is not so clear that we should ignore degrees of belief of the observer.

Assume that any subspace $\mathcal{H}_\mathcal{S}$ of a given Hilbert space $\mathcal{H}$ defines an observable:

$$[\psi_S] = |\psi_S\rangle \langle\psi_S| \tag{3.116}$$

where the application of this observable to the wavefunction $|\psi\rangle$,

$$[\psi_S]\,\psi\rangle = |\psi_S\rangle \langle\psi_S|\,\psi\rangle \tag{3.117}$$

essentially asks $|\psi\rangle$ whether it has state $|\psi_S\rangle$, by projecting it onto the subspace $\mathcal{H}_\mathcal{S}$ defined by $|\psi_S\rangle$. Now suppose we have two different sets of mutually orthogonal projectors that completely define

[34] The original proof, which was based on PVMs, made an additional assumption that the dimension of the Hilbert subspace was greater than 2. However, this limitation has been eliminated in more modern POVM-based versions of the proof [44].

the possible outcomes of two different measurements, $\{[i]\}$ and $\{[i']\}$, but which both *share* one of their outcomes—say the j-th one—so that

$$[j] = [j'] \tag{3.118}$$

Assumption 3.26. ***Gleason Noncontextuality:*** *If* $[j] = [j']$, then $p([j]) = p([j'])$ independent of the bases that $|j\rangle$ and $|j'\rangle$ each belong to.

Since $|j\rangle$ and $|j'\rangle$ are analytically the *same* state, noncontextuality thus means that the probability of an outcome is dependent only on the state, and is independent of "context" (*i.e.,* of which basis the measurement places that state in).[35]

Even without assuming noncontextuality, additivity and summation to unity allow us to group all the alternative outcomes in both cases as the complement of the outcome in question [135], so that:

$$\begin{aligned} p([j]) + p([\bar{j}]) &= 1 \\ p([j']) + p([\bar{j}']) &= 1 \end{aligned} \tag{3.119}$$

However, since $[j] = [j']$, it follows that $[\bar{j}] = [\bar{j}']$ (Th. 2.8). Thus, both of the above summations are functions of the *same* projectors, and noncontextuality follows:

$$p([j]) = p([j']) \tag{3.120}$$

so long as probabilities are functions *only* of the associated projectors. But for probabilities to rely on anything else would require that they rely on something other than the actual analytic state of the system in question—and assuming wavefunction realism, this means something other than the physical state. Thus, noncontextuality seems to be equivalent (in an Everettian context) to simply assuming that probabilities rely only on physical states, which is what Wallace calls "state supervenience". [226].

It would appear, then, that any view of probabilities that associates a probability with a quantum state, so that the probability of state $|\varphi\rangle$ obtaining after unitary transformation,

$$|\psi\rangle \Rightarrow |\psi'\rangle \rightarrow |\varphi\rangle \tag{3.121}$$

[35] Note that Gleason noncontextuality is completely different than Kochen-Specker noncontextuality, which says that the state of the measured variable is independent of the measuring situation (so the "context" is the actual physical or analytic measuring situation, not the synthetic choice of basis). This type of noncontextuality was shown to be incompatible with hidden variables theories in [118], and thus has a similar import to Bell's theorem. It is generally assumed that quantum measurements are *synthetically noncontextual* (Gleason's sense) but *analytically contextual* (Kochen-Specker's sense). To avoid confusion, we can call Gleason's contextuality "synthetic contextuality", and Kochen-Specker's version "analytic" or "physical contextuality".

is a function $f()$ only of the initial (pre-measurement) state and final (outcome) states:

$$p(|\varphi\rangle) = f(|\varphi\rangle, |\psi\rangle) \tag{3.122}$$

ought to be able to simply cite Gleason to prove the Born rule.

3.3.12.2 Gleason's Theorem (POVM version)

Once we accept non-contextuality, Gleason's theorem is a straightforward formal proof about measures on Hilbert spaces. In its original form [91], it was about projective measurements, was notoriously difficult, and was restricted to dimensions of 3 or greater. However, it has since been generalized to POVM measurements [38, 44], which greatly simplifies the proof, even removing the dimensionality restriction.

Recall from §2.5.4 that a POVM is defined as a set of potential effects $\{\lceil i \rceil\}$ that decompose identity, $\sum_i \lceil i \rceil = \hat{I}$, with eigenvalues in $[0 \cdots 1]$. Define $\mathcal{E}(\mathcal{H})$ as the set of all potential effects in Hilbert space $\mathcal{H}$.

We are interested in the possible forms that a probability measure $p(\lceil k \rceil \,|\, \{\lceil i \rceil\})$ on $\mathcal{E}(\mathcal{H})$, obeying additivity and summation to unity, could take.

Theorem 3.27. ***Gleason's Theorem (POVM version).*** *Any probability measure $p(\lceil k \rceil)$ on $\mathcal{E}(\mathcal{H})$—obeying additivity and summation to unity—is of the form $p(\lceil k \rceil) = \mathrm{Tr}\left([\rho]\,\lceil k \rceil\right)$, for some density operator $[\rho]$.*

Proof. The following largely follows [38]. Clearly, for all natural numbers $n \geq 1$, and for any $\lceil k \rceil$,

$$p(\lceil k \rceil) = np(\frac{1}{n}\lceil k \rceil) \tag{3.123}$$

$$\frac{1}{n}p(\lceil k \rceil) = p(\frac{1}{n}\lceil k \rceil) \tag{3.124}$$

Setting $x = 1/n$, we have, for all rational fractions x in $(0 \cdots 1]$:

$$xp(\lceil k \rceil) = p(x\lceil k \rceil) \tag{3.125}$$

Additivity and positivity mean that any measure on $\mathcal{E}(\mathcal{H})$ is order preserving:

$$\lceil k \rceil \leq \lceil j \rceil \rightarrow p(\lceil k \rceil) \leq p(\lceil j \rceil) \tag{3.126}$$

Let $\{a_i\}$ and $\{b_j\}$ be sequences of rational numbers in $[0 \cdots 1]$ such that $\{a_i\}$ approaches x from above and $\{b_j\}$ from below:

$$\{b_j\} \nearrow x \swarrow \{a_i\} \tag{3.127}$$

It follows that for any $a_i \in \{a_i\}$ and $b_j \in \{b_j\}$:

$$\begin{aligned} p(b_j \lceil k \rceil) &= b_j p(\lceil k \rceil) \leq p(x \lceil k \rceil) \leq p(a_i \lceil k \rceil) = a_i p(\lceil k \rceil) \\ p(b_j \lceil k \rceil) &\leq a_i p(\lceil k \rceil) \end{aligned} \tag{3.128}$$

Let $\hat{A}$ be any positive bounded operator so that $\hat{A} \notin \mathcal{E}(\mathcal{H})$. This means there is some $n \geq 1$ and $\lceil k \rceil \in \mathcal{E}(\mathcal{H})$ such that

$$\hat{A} = n \lceil k \rceil \tag{3.129}$$

Assume we have $\lceil i \rceil, \lceil j \rceil \in \mathcal{E}(\mathcal{H})$ such that

$$\hat{A} = n_1 \lceil i \rceil = n_2 \lceil j \rceil \tag{3.130}$$

and assume (without loss of generality) that $1 \leq n_1 < n_2$. Then

$$\begin{aligned} p(\lceil j \rceil) &= \frac{n_1}{n_2} p(\lceil i \rceil) \\ n_2 p(\lceil j \rceil) &= n_1 p(\lceil i \rceil) \end{aligned} \tag{3.131}$$

And we can uniquely define

$$p(\hat{A}) = n_1 p(\lceil i \rceil) \tag{3.132}$$

Now let $\hat{A}$,$\hat{B}$ be positive bounded operators. Consider $n > 1$ such that

$$\frac{1}{n}(\hat{A} + \hat{B}) \in \mathcal{E}(\mathcal{H}) \tag{3.133}$$

Then we have

$$\begin{aligned} p(\hat{A} + \hat{B}) &= np\left(\frac{1}{n}\left(\hat{A} + \hat{B}\right)\right) \\ &= np\left(\frac{1}{n}\hat{A}\right) + np\left(\frac{1}{n}\hat{B}\right) \\ &= p(\hat{A}) + p(\hat{B}) \end{aligned} \tag{3.134}$$

Finally, let $\hat{C}$ be an arbitrary bounded Hermitian operator. Assume we have two different decompositions:

$$\begin{aligned} \hat{C} &= \hat{A} - \hat{B} \\ &= \hat{A}' - \hat{B}' \end{aligned} \tag{3.135}$$

We therefore have

$$\begin{aligned} p(\hat{A}) + p(\hat{B}') &= p(\hat{B}) + p(\hat{A}') \\ p(\hat{A}) - p(\hat{B}) &= p(\hat{A}') - p(\hat{B}') \end{aligned} \tag{3.136}$$

Thus we can uniquely define

$$p(\hat{C}) = p\left(\hat{A} - \hat{B}\right) = p(\hat{A}) - p(\hat{B}) \tag{3.137}$$

It is then straightforward to extend this linearity to the general case, and it was already well-established [223][63, p.6, Lm.1.6.1] prior to Gleason that any generalized probability measure $p()$ on potential effects extends to a unique positive linear functional, which is normal and obtained from a density operator $[\rho]$, so that $p()$ has the form

$$p(\lceil k \rceil) = \mathrm{Tr}\left([\rho]\,\lceil k \rceil\right) \tag{3.138}$$

which is the Born rule. □

3.3.12.3 Discussion

Gleason's proof is often seen as unrelated to Everett's, but it is actually closely related to Everett's stage 1, since Gleason's proof (in its projective version, at least [135]) can be executed by first proving amplitude dependence from noncontextuality. Hence, the remainder of the proof can more or less be replaced with Everett's stage 1, which assumes amplitude dependence. Both proofs are essentially proofs of linearity from additivity.

Since Wallace [226] assumes state supervenience (which implies noncontextuality) and derives amplitude dependence from there, we could presumably also replace much of Deutsch-Wallace with Everett's stage 1. Hence, Gleason, Everett (stage 1) and Deutsch-Wallace are all closely related proofs, that try to prove the Born rule to be the only measure applicable to quantum states—in opposition to frequentist proofs which are focussed on reducing the Born rule to a completely different rule, which is given synthetic *a priori* validity.

Note, however, that Gleason's assumptions are weaker than either Everett's or Wallace's, and his result correspondingly stronger. Hence, Gleason's proof seems tailor-made to serve as an MWI Born rule proof. It is highly analytic, showing that the Born rule follows almost from the formal structure of the wavefunction alone, with the only synthetic axiom being very likely benign, at least from an objectivist stance. And given that Everett's postulate of wavefunction realism assumes that the wavefunction state is the full ontology of the system, noncontextuality is likely to be a benign assumption for MWI, so long as we accept the idea of objective probabilities (and I will argue in more detail, in Ch. 8, why I think this is the case for ASU). This would seem to imply that our view of the interpretation of probability theory will determine whether we consider noncontextuality to be acceptable. Hence, it would seem that Gleason's theorem should be at least highly relevant to the MWI probability debate, even if it does not settle the matter definitively.

Yet, curiously, Gleason's theorem has not been cited much in the Everettian probability literature, and is not even generally seen as relevant [238] by either side, perhaps in part because its assumptions and approach are much different than Everett's. But this may be nothing more than an historical accident. Gleason's proof may have overlooked potential, if the assumption of noncontextuality can be shown to be a workable (synthetic) *a priori,* given the existence of conscious mechanical subsystems in the wavefunction—especially if we have already rejected branch-counting as a valid *a priori*. One of the reasons, I believe, that the Born rule objection survives the existence of Gleason's theorem is that branch-counting is considered to provide a kind of *reductio ad absurdum* for the MWI. It is not so much that Gleason's theorem is thought to not apply to the MWI, but that the MWI, by its own lights, (supposedly) demands branch-counting, and hence already does *not* follow the Born rule.

It also seems likely that the idea of Gleason as irrelevant arose partly out of the frequentist belief that infinite measurement sequences *were* relevant, *a priori*. Thus, the derivation of a mathematically sound measure for single cases (even if we can show from reasonable assumptions that this is the *only* possible measure) came to be considered, at best, a starting point, since it still conflicts with the world-counting statistic, which was considered a mandatory *a priori*—and infinite sequences are the frequentists' way of answering this world-counting challenge.

However, once we reject this whole set of assumptions (which I have done already, to some extent, and will continue to do through Ch. 4), we are no longer tied to frequentism, world-counting is no longer an *a priori*, and there is no reason not to seek a single-case measure.

As for noncontextuality, it would seem that this assumption ought to be allowed in an *objective* account of probabilities, such as I will be advocating for in Ch. 4, where we can replace the human measurer with a bunny rabbit that has no conception of what it is measuring (or that any measuring is even happening), and the same probability for any given outcome should hold. It certainly seems that if probabilities are to be objective in this sense, they must rely only on the actual analytic or physical states in question. Branch-counting, therefore, will not serve as a probability measure for bunny rabbits, by Gleason's theorem. It *might* serve for humans, if by "probability" we mean something subjective—something fundamentally inapplicable to bunny rabbits—and dependent on the peculiar beliefs of humans about what is going on (and, of course, I am not discounting that subjectivists may have their own justifications for noncontextuality).

I will argue in Ch. 8 that, in the context of ASU—which I will further develop throughout the coming chapters—the possibility of contextuality simply does not arise. Hence, Gleason may well provide an adequate proof of the Born rule in this context. Although Born rule objectors may still be able to produce *reductio* arguments, these will have to actually demonstrate inconsistency in the

algorithmic interpretation. This will be difficult to do, I think, without an alternative *a priori* rule to put forward, and I will attempt to show that the favorite choice (branch-counting) is by no means required—and even contraindicated—by an algorithmic approach to the MWI.

Thus, counter to common wisdom, I consider Gleason quite relevant to the MWI Born rule debate, much more so than the frequentist proofs. In fact, the best approach to a proof of the Born rule for an algorithmic MWI may not be the construction of a brand new Born rule proof at all, but rather simply a defence of the logical coherence and simplicity of the algorithmic approach, since the very assumptions of this approach do not, according to Gleason, allow for any probability rule *but* the Born rule. Nonetheless, this is no small task, as the assumptions of my approach are not uncontroversial, probably not even amongst those who would consider themselves computationalists and/or Everettians. The defence of the coherence of ASU will be my task in Ch.6-8, which will conclude with a re-statement of my claim of the adequacy of Gleason's theorem in this context.

A final note on Gleason: according to [200], the no-signalling theorem can be considered a consequence of noncontextuality. This theorem asserts the impossibility of using quantum entanglement between subsystems to communicate messages between the subsystems, implying that quantum entanglement does *not* (as it may at first blush appear to) violate Einstein's prohibition on faster-than-light signals. This raises the possibility of deriving the Born rule from the synthetic *a priori* impossibility of faster-than-light signalling. An argument could possibly be made that a world with such instantaneous signalling is impossible, perhaps because it would be incompatible with regular causality and the arrow of time (*i.e.* it would imply causality from the future to the past). A world that permitted such backwards causation might be incompatible with conscious observers. One would have to show, however, that *any* (or very nearly any) degree of such superluminal influence would be inconsistent with consciousness. A proof merely that an inordinate amount of it would "mangle" conscious observers would not in itself be enough (so merely invoking the grandfather paradox[36] is only a start).

3.3.13 Do Maverick Worlds Count?

Central to the debate over Everett's stage 2 proof is the presence in the wavefunction of "maverick worlds", meaning very large non-Born measurement sequences, whose amplitudes approach zero in the limit. These worlds are a problem because they are still there in the wave function, given that we do not actually ever reach the fabled infinite sequence with zero amplitude. But even once we have

[36] The grandfather paradox is the classical argument against the possibility of time-travel (the most blatant kind of backwards causation). If time-travel were possible, and I travelled back in time and killed my grandfather before my parent's were even born, what would become of me?

rejected world-counting, and offered a plausible algorithmic alternative, maverick worlds are still an interesting philosophical issue for Everettians to face. Some people consider the mere presence of such worlds, even of of low probability, a thorny issue.

We first need to distinguish between two kinds of mavericks: maverick worlds that are clearly analytically in the wavefunction at low amplitude, and those that are merely possible worlds, not in the wavefunction. It is by no means clear that these are two separate things, but not everyone is going to equate them, so we need to make the distinction. Take, for example, the classic EPR thought experiment. We said in Ch. 2 that Alice and Bob must necessarily measure opposite spins. This would seem to indicate that the possible world in which Alice and Bob both measure "+" (undeniably, this is a *possible* world) actually has zero probability in the wavefunction. Hence "possible world" must be a larger set than "worlds represented in the wavefunction of the universe."

However, this does not follow. We have already suggested that even objective probabilities may require a synthetic element. From this point of view, a probability is fundamentally a probability *for* a perceptual experience, not for just any physical event or analytic structure. Thus, the expression $p(+)$, under this view of probabilities, should be read as "the probability of perceiving a + on the measuring apparatus". Whether or not there *really is* a + reading on the apparatus is not pertinent to the probability. In fact, under this view, we must assume that contributing to the absolutely precise value of $p(\text{Alice}_+ \ \& \ \text{Bob}_+)$ are extremely tiny amplitude worlds in which Alice, say, hallucinates that she sees a + reading, or where the molecules in the pointer device of the measuring apparatus just happen to move the pointer to the + position. In other words, *no* world, defined perceptually (which is how they must be defined for Everett) is completely excluded from the wavefunction. Hence, the set of worlds represented in the wavefunction really does include all possible worlds.

Of course, even if one argues that the world in which Alice hallucinates a + result should *not* be included in the wavefunction under Alice_+, it is nonetheless still in the wavefunction! So there remains an amplitude for Alice to experience seeing a + result, whether we like it or not, and whether or not we define it to be part of what we mean by Alice_+. All kinds of maverick worlds with all kinds of ridiculous things happening *are* in there, likely including any possible continuation of one's experience that one could have.

On the face of it, there is no inherent reason why the mere presence of these worlds should be a problem, so long as their probabilities are vanishingly low, and we have not invoked world-counting. But, again, their mere presence makes some feel uncomfortable. For instance, it is hard to see why this would not include worlds where, for instance, the molecules in the air around you suddenly self-organize into a magical green elf. Everett's interpretation would seem to require you to admit that this magical world "really exists". On the other hand, so what? Given the Born rule, there is not

necessarily a problem here: the green elf world, let's call it Elfworld, has extremely low amplitude, and therefore extremely low probability. Why should it bother us that it exists, if its probability is so tiny that the chance of ending up in it is negligible? Even more problematic than the existence of Elfworld is the debate over whether the presence of maverick worlds implies some kind of universal immortality [146, 215, 172, 155, 138, 214, 174], given that, once one has died in all the Born worlds, the maverick worlds are all that is left, and so they are maverick no longer. Indeed, Everett himself, seems to have believed something like this [40].

While I do not myself see a problem with extremely low-probability maverick worlds, we will return to this issue in Ch. 6, and I will argue that, in fact, such worlds, *qua* worlds, do not really exist at all, due to the synthetic *a priori* nature of what we mean by "world". My argument is similar to Hanson's idea [102] of low-amplitude worlds being "mangled" out of existence by high-amplitude ones. It might seem that the immortality arguments escape refutation via world-mangling, given that the mavericks are the only worlds left after one dies in all the non-maverick worlds. However, I think there are good arguments [138, 174] why the idea of quantum immortality is fundamentally flawed.

Of course, the biggest problem with maverick worlds comes from the fact that, for so many, the idea of world-counting (or observer-counting, or outcome-counting) seems self-evident. I suspect many will still believe in world-counting, even if they accept that it cannot be proven, at least not analytically. An argument could possibly be made that world-counting is justified as a synthetic *a priori*, but such an argument would imply a lot of further things about one's interpretation of probabilities, which we will examine at more length in the next two chapters.

For now, we will look briefly at the basic idea that probabilities should be based on worlds or observers. As we have seen, there are basically two versions of the Everett interpretation:

1. *Ontic branch version*: in this version, the branches are taken to be the ontic entities that make up the stuff of the wavefunction. This can be further subdivided according to whether one views the branches as being fundamentally:
   (a) *worlds*,
   (b) *observers*, or
   (c) *outcomes*.
2. *Pure wavefunction version*: this version sticks strictly to wavefunction realism, so only the wavefunction of the universe is ontic, and worlds are merely emergent.

The ontic branchers will clearly have a predilection for counting branches. However, they have the problem of defending why the ontic entities of the wavefunction seem to be, from an analytic perspective, completely emergent and optional features of it. The pure wavefunction Everettians are being truer to Everett's postulates, but are going against the assumptions of his stage 2 proof. They also have the problem, if they believe in count-based probabilities, of explaining what to count in an

ontology with only a single existing thing. Some pure wavefunction advocates still end up counting branches, but it would seem that this either puts them in the camp of the subjectivists, or else puts an onus on them to explain why they should be counting emergent branches, if branches are not ontic entities.

So is branch-counting justified? And, if so, what kind of branches are we counting: worlds, outcomes or observers? Recall the marble example from earlier. Here, we presumed that we should not count observers, but marbles, since it was marbles that were the actual ontic entities. But mightn't it be argued that we *should* count observers? So what if Observer A emerges from 100 marble picks, and Observer B from only one? Since Observer A is identical in all those 100 picks, there is arguably still only one observer, not 100. From *that* observer's point of view these are not 100 states in superposition, but only one state. So perhaps it should only count as one. On the other hand, we could decide to go ahead and count "100", and ignore the straight observer count.

One of the problems here is that probability theory does not really work as a "law". It only tells us what will "probably" happen. It then becomes hard to *prove* any statistical rule for what to count. In the real world, we can use common physical intuition to see what needs to be counted. This fails us for the quantum wavefunction, as it does not correspond to our natural physical intuitions. But we can at least get a start on these issues by applying our common physical intuition by analogy to something more physically familiar, and see if this helps. Imagine, instead of Everettian worlds, that we have an ensemble of "worlds"—but in the sense of *planets* within the same universe[37], not isolated spatio-temporal universes. Assume there are 100 such worlds circling 100 suns, all much too far away from each other for any hope of communication. Label them $\{w_1 \cdots w_{100}\}$. Assume these are the only inhabited worlds in the universe (and that there is only one universe).

We will ask two questions about probability rules for these worlds, one with respect to global anthropic constraints, and the other with respect to events within the worlds:

1. Anthropic Constraints: Presume that w_1 has a population of $1,000,000$ and a primarily *oxygen* atmosphere, while all the rest of the worlds have a population of only 1, and a primarily *methane* atmosphere. What is the likely *a priori* state of the world: should a random observer expect to be breathing oxygen or methane? Well, if we count worlds, then we should expect methane. But if we count observers, we should expect oxygen. But here, physical intuition and common sense strongly tell us that we should count observers. What matters is that, for the vast majority of conscious observers in the universe, they are breathing oxygen. So if we consider ourselves a random pick out of conscious observers—arguably what the anthropic principle demands—then surely we should be counting observers and not worlds.

2. Branching: now imagine that all 100 worlds each have one observer on them. Imagine further that the number of degrees of freedom in these worlds are small enough that it is feasible for all the worlds to end up with observers that are all in the identical conscious state $\mathcal{C}$ at time t_1.

[37] The basic form of this thought experiment I get from Jacques Mallah.

> This is just a coincidence, not a result of any communication between the worlds. However, there *are* some tiny differences between the worlds, which cause these identical conscious states to bifurcate at time t_2, so that we have a new conscious state C_1 in $\{w_1 \cdots w_{25}\}$, and C_2 in $\{w_{26} \cdots w_{100}\}$. Now we have two observers—or at least two distinguishable observer states—but still 100 "ontic" worlds. If we accept the Strong AI postulate, then there really are only two observers. The one observer has essentially "split" at time t_2 into two different observers. We will be agnostic here as to whether there "really are" only two consciousnesses or observers here, or whether there "really are" 100 and some just happen to be identical. Now the question is: while remaining agnostic about the "real" number of observers, can we say what the probability is for any given observer going from t_1 to t_2? Will they see C_1 with probability of 25%, because it happens in 25% of the worlds? Or will they see it with probability 50%, because there are effectively only two observers? It seems, again, that our physical intuition and common sense tells us that clearly the probabilities are 25%/75%, not 50%/50%. This ought to be accepted by the Strong AI advocate who actually believes the set of twenty-five $\mathcal{C}_2$ observers is one conscious person, *and* by the person who believes they are twenty-five distinct but identically conscious persons.

Applying this intuition to the wavefunction, we see that whether we believe that separate ontic entities yielding identical consciousnesses are one conscious person or distinct persons, we should still accept that we need to count ontic entities not observers. Argument *(1)* above makes it seem intuitive that observers take precedence over worlds, but part *(2)* shows that it is not really worlds versus observers that is the underlying issue, but ontic entities versus observers.

It would seem, then, that we are left with counting some kind of analytic ontic entities, versus counting something wholly synthetic. If the former, we can make a good case for amplitudes, but then it is difficult to quite go along with some of Everett's language, about the wavefunction being the only real thing in the universe. If there is only one real thing, we have nothing *a priori* to count anymore (not if there is only one of them). To make the leap from counting wavefunctions to counting amplitudes would require an understanding of how it is that amplitudes are something it makes sense to count.

Further discussion of these issues will have to wait, as none of this can be resolved any further unless we decide what we think probabilities actually *are*, and this is a matter of some controversy. In the next chapter, we will examine some of the main issues in the foundations of probability theory, and we will end up coming to the conclusion that we need to count *analytic* ontic entities, whereas we need to *categorize* these entities into *unitary-synthetic* equivalence classes.

4 Probability Theory

4.1 Measure Theory

Probability is a particular kind of mathematical "measure", which is any function that tells us "how big" something is. Specifically, in terms of sets:

Definition 4.1. A *measure* on a set Ω is a function $\mathcal{M}(\varepsilon \subseteq \Omega)$ that, given a subset ε of Ω, returns a number that meets the following criteria:

1. *Non-negativity*: $\forall \varepsilon \subseteq \Omega : M(\varepsilon) \geq 0$ (since what would "negative size" mean?)
2. *Countable additivity*:
   $\forall (\mathcal{E} = \{\varepsilon_k : \varepsilon_k \subseteq \Omega\}; \varepsilon_i, \varepsilon_j \in \mathcal{E}; \omega_q \in \varepsilon_i, \omega_r \in \varepsilon_j : \omega_q \neq \omega_r) : \mathcal{M}\left(\bigcup_k (\varepsilon_k \in \mathcal{E})\right) = \sum_k \mathcal{M}(\varepsilon_k \in \mathcal{E})$

Countable additivity means that, for all sets $\mathcal{E}$ of subsets of Ω, if all members of $\mathcal{E}$ are *disjoint* (having no members in common), then the measure on the union of all members of $\mathcal{E}$ is the same as the sum of the measures on the members of $\mathcal{E}$.

The reason we require countable additivity is simple: Set $\mathcal{E}$ is a collection of subsets of the main set that our measure is on. If no two of these have any members in common, then there is no overlap between the subsets: each one represents a single "piece" of the whole (Ω). So if we measure the size of a collection of such pieces, it had better be the same as the sum of the measure of all the individual pieces. If I cut a piece of bread into three (equal or unequal) pieces, the volume of the piece of bread will equal the sum of the volumes of the three pieces. In terms of probability theory, if the probability of rolling a 1 on a die is $1/6$, and the probability of rolling a 2 is also $1/6$, then the probability of rolling a 1 *or* 2 will be the sum, $1/6 + 1/6 = 1/3$, so long as the two outcomes are *independent* of one another.

In discrete probability theory, the requirement of countable additivity does not tend to be too problematic. It is in a continuous domain that it creates problems. For instance, we would normally think that, if a spatial region or object were divided into several sets of non-overlapping pieces, that the sum of the volumes, or Lebesgue measures, of these pieces will have to be equal to the measure of the whole. In fact, however, for some sets, countable additivity does *not* hold for the Lebesgue measure (and we say that such sets are not Lebesgue measurable).

These problems only arise for continuous spaces, however, so we can ignore them here. For us, the volume measure does not have such problems, since we are permitted to speak of measures only on discrete spaces, to begin with. Limit expressions may be invoked, so that we may say that the limit of a measure goes to X, as the space becomes more and more finely-grained, but we cannot turn this into an actual measure on a continuous space.

4.2 Probability Theory

Probabilities are a kind of measure: the kind that measures the chances of some event (or outcome or situation) happening. Although, in common language, "outcome" and "event" may seem like pretty much the same thing, in probability theory, an "outcome" is an actual individual result, while an "event" is a set of such outcomes, or "equivalence class", whose members we decide to consider effectively equivalent, for the purpose of calculating probabilities. In statistical mechanics, the outcomes are called "microstates", and the events are called "macrostates".

Definition 4.2. A *probability measure* on a *sample space* Ω is a measure on that space, such that

1. The *sample space* Ω is the set of all *outcomes* or *microstates* (individual things that are possible),
2. An *event* or *macrostate*, $\varepsilon \subseteq \Omega$, is a subset of sample space Ω (grouping together outcomes we want to count as the same),
3. An *event* or *macrostate set* $E \subseteq 2^{\Omega}$ is a subset of the set of all possible events (specifying which events we wish to consider),
4. The *probability measure* $p(\varepsilon \subseteq \Omega)$ satisfies the *normalization criterion*:

$$p(\Omega) = 1 \tag{4.1}$$

The reason for #4 is obvious: if we treat all possible outcomes the same, grouping them together into one event, then the probability of this all-encompassing event *must* be unity, since it is guaranteed to happen.

Given countable additivity, it follows that for any event set where the events are disjoint (mutually exclusive or independent of each other), then the probability of any subset of the event set will equal the sum of the probabilities of its events:

$$\forall \,\text{disjoint}\, \mathcal{E} \in E:\ p\left(\bigcup_k (\varepsilon_k \in \mathcal{E})\right) = \sum_k p(\varepsilon_k \in \mathcal{E}) \tag{4.2}$$

In other words, the probability of the conjunction of any number of mutually exclusive events must equal the sum of the probabilities of all the individual events. Combined with the normalization

criterion, this tells us that the sum of all events in an event set must be unity:

$$\sum_k p\left(\varepsilon_k\right) = 1 \tag{4.3}$$

Given this, it also follows that the individual outcomes have probabilities that also sum to unity:

$$\sum_k p\left(\omega_k \in \Omega\right) = 1 \tag{4.4}$$

where we allow that each outcome $\omega_k \in \Omega$ can also be considered an event with only one member:

$$p\left(\omega_k\right) = p\left(\{\omega_k\}\right) \tag{4.5}$$

We can now give a general definition for the probability of an event, in terms of the probabilities of its outcomes.

Definition 4.3. The *probability* of an event ε is

$$p\left(\varepsilon\right) = \sum_k p\left(\omega_k \in \varepsilon\right) \tag{4.6}$$

In other words, the probability of a particular event is just the sum of the probabilities of its outcomes. This completes the basic notion of a probability: it is a measure with total measure unity that is non-negative and obeys countable additivity. Without these features, it would not behave like a probability.

Note, however, that our definition of probability does not tell us what probability to assign to *individual outcomes*. These are the primitives of our system, and so they must be assigned prior probabilities from outside the main postulates of probability theory.

4.3 The Interpretation of Probability Theory

Probability theory itself is merely a mathematical framework. It tells us nothing about how to apply this mathematical machinery. The application of probability has given rise to various controversies on its application, yielding different *interpretations* of probability.

Because probability theory makes reference to primitive probabilities that must be assigned a prior value from outside the theory, it can be argued that the mathematical theory of probability does not tell us what the word *probability* even means in a more general context. This is why it is sometimes said that the meaning of "probability" is actually not a purely mathematical matter, but belongs more properly in the realm of the philosophical, perhaps even the metaphysical.

Of course, we could always argue for a minimalist interpretation, claiming that "probability" has whatever meaning it gets from the mathematical framework alone, and whatever is then left unsaid

is a matter for individual applications. However, this will still leave us with competing schools of thought on the best general guidelines for how to go about applying probability—so either way, we are still stuck with the competing interpretations.

In this dissertation, we are dealing with the Born rule objection, and we have already seen that certain assumptions about what one should count—worlds, observers, amplitudes, etc.—result in completely different views on the adequacy of the Everett interpretation. But how can we determine what entities we should count to calculate probabilities—or even if we should be counting entities at all—if we don't have any clue what we mean by "probability" in the first place (or, equivalently, if we have no general guidelines for how to apply probability)?

So we need to first take a look at the major competing interpretations of probability theory, if we are to have any hope of settling these conceptual problems. A complete analysis of all the competing interpretations, and their pros and cons, is well beyond the scope of this dissertation. For our purposes, I will focus on finding a view of probability that suits application to an Everettian context, but will not presume to solve the entire general problem of interpreting probability theory, nor even to answer all the possible objections to the view of probability I will adopt.

4.3.1 Kinds of Probability

Before looking at individual interpretations, it will help to distinguish between several kinds of probability. I will categorize probabilities in four different ways, according to these distinctions:

1. *Objective versus subjective:* probabilities can be placed on a spectrum between "objective" and "subjective". Objective probabilities take on their specific values independently of what we think about them—including everything we know, believe or feel about them (and even independently of whether we know, believe or feel *anything* about them at all). Subjective probabilities are dependent, at least in part, on what we think about them (and are hence not even possible unless we think *something* about them).

2. *Discrete versus continuous*: discrete probabilities are computed strictly from discrete values with finite information content. Continuous probabilities are computed (or imagined to be computed) from continuous values that contain an infinite amount of information.

3. *Single-cases versus ensembles*: probabilities can be applied to (or derived from) either single-cases (where no reference to other situations or experimental setups is required or implied), or ensembles of cases.

4. *Causal versus statistical*: probabilities can be viewed either as deriving from underlying causal forces in the world (or some other ontology), or merely as descriptions of the resulting observed statistics.

One's choice of interpretation of probability theory will often flow from how one prioritizes these distinctions. The first distinction above—objective versus subjective—has been given perhaps an

undue amount of the attention in the probability interpretation debate. While it is an important distinction, it is too often presumed that once one decides where one stands on this issue, the rest simply falls into place. In my opinion, this is anything but the case, and all four of these distinctions are important.

4.3.1.1 Objective versus Subjective

Three kinds of probability It is not uncommon to simply divide this spectrum into the two types "objective" and "subjective". However, I will follow Mellor [142] in maintaining that the use of only two categories blurs important distinctions and can cause confusion. The three types I will use are as follows (this largely follows Mellor's approach, but the terminology and usage are not precisely the same):

1. *Ontic probabilities (*or *chances)* are unique probabilities that would result from complete knowledge of the world; in other words, probabilities that are completely objective and independent of the state of our knowledge or beliefs about the world.
2. *Epistemic probabilities* are unique probabilities that are a result of partial knowledge of the world.
3. *Doxastic probabilities (*or *credences*) are probabilities that are (*qua* credences) unrelated to knowledge of the world, representing degrees of subjective belief.

A quantum probability is an example of a probability that is likely ontic, if anything is, since it seems to be mandated by physical theory, not one's state of knowledge. A person is still subject to the same quantum probabilities, whether one knows anything about the world or not.

The probability of the truth of some theory, based on available evidence, would be an example of an epistemic probability. For instance, the probability that string theory is correct, or that the butler did it. Clearly, string theory is either correct or incorrect; and the butler is either guilty or innocent. If we say that one of these theories has a probability of 38% of being correct, this is clearly a judgement of the quality of our evidence, not an objective fact about the world.

A doxastic probability is simply a degree of subjective belief that is subject to the mathematical rules of probability theory. Such belief may or may not be justified by evidence, and may or may not be objectively true of the world. So a credence might *also* be an epistemic probability or a chance; but it is, *qua* credence, unrelated to such epistemological and metaphysical considerations. An example of a credence that is not a chance or epistemic probability might be your degree of belief in your spouse when he tells you he is not cheating on you. While it is feasible that this might be based on evidence, it is also possible that it is based solely on gut intuition, or wishful thinking, with no clear discernible evidence available for it, one way or the other.

Three Interpretations An argument could be made for any one of these types to be considered the most fundamental sense of "probability", with the other two being merely special, idealized or degenerate cases of it. Let's look briefly at these three possible types of interpretations of probability:

1. *Metaphysical interpretations*: One could argue that probabilities are fundamentally *ontic* or metaphysical, referring to a reality independent of our knowledge, while epistemic and doxastic probabilities are simply our (often inadequate) attempts to capture this reality with our limited resources. After all, whatever the limited state of our knowledge, it is still knowledge *of* the world. Whatever degree of uncertainty we may have in our belief, it is always belief *about* the world. So we are always ultimately trying to make our assigned probabilities match up with the true objective probabilities. Thus, probabilities are fundamentally ontic in nature.

2. *Epistemological interpretations*: Or, one might argue that it is the notion of an *epistemic* probability that is most basic, while purely ontic and doxastic types are simply the extreme idealized ends of the epistemic spectrum, allowing for the unlikely possibility of complete knowledge on one end, and the equally unlikely possibility of belief with no knowledge on the other. Thus, while we can idealize about the degenerate cases, probabilities are most centrally epistemic in nature.

3. *Psychological interpretations*: Finally, one might instead argue that probabilities are basically doxastic, so that credences are most fundamental. After all, when we look to find probabilities, we are always concerned with establishing or discovering credences. While we can have probability without evidence or world knowledge, wherever there is probability, there is always credence. Hence, probabilities are ultimately doxastic in nature.

A fourth kind It can be useful to informally create a fourth level of objectivity that straddles the divide between the ontic and epistemic. I have cited scientific theories, with their supporting evidence, as typical of the kind of statements to which we assign epistemic probabilities. However, what about a classical coin flip? (We will assume here that we can set up a coin toss in such a way that the outcome has no dependence on quantum randomness.)

Most people would find it odd to say that a classical coin toss yields a probability based on "evidence". Instead, we would rather think of the probability as (sort of) objective in nature. In other words, we think of coin tosses more in terms of chances—*as if* they really were chances, and not epistemic in nature (even though we know that they are, strictly speaking, epistemic). Clearly, they *are* epistemic, since if we knew *everything* relevant about the setup of the toss, right down to all the relevant muscle movements, molecular configurations and forces, we could always assign a probability of either 0 or 1 to any particular toss of the coin.

However, most of us would find it awkward to describe our partial knowledge of a coin toss setup as "evidence" for or against an outcome of heads. Instead, we think of our model of the situation as more or less adequate, but with some inherent randomness. This is because we do not really have a desire to go around forming theories, and collecting evidence, for or against the outcomes of individual coin tosses, as this would be a frightful waste of our time. The knowledge of the world

we lack here is not *important* knowledge. It is knowledge we are happy to overlook. So, instead of viewing this uncertainty as a property of a theory, we see it as an objective property of a coarse-grained model of the world, which we treat for practical purposes almost as if it were a complete model.

This is a matter of degree, not of kind, since for both theories and coarse-grained models, the probabilities still depend on our partial knowledge of the world, and so the probabilities are still strictly epistemic in both cases. One *could* conceive of the uncertainty over string theory as a (very) coarse-graining of our world model, just as one *could* conceive of the uncertainty over heads or tails as a matter of two competing theories with equal evidence supporting them. However, we tend to think of string theory, and the butler's guilt or innocence, as matters of evidence for theories; while we tend to think of coin flips as matters of (almost) objective chances.

And while this tendency is understandable, even useful, it can create confusion, since it can lead us into the temptation of calling a classical coin toss probability a "chance", and then assuming, explicitly or implicitly, that it is entirely objective. Hence, I will allow that when a probability is *technically* epistemic, but relies on a well-established but coarse-grained model, that we may informally call it an *epistemic chance* (technically, a contradiction by our above definitions, but still a useful term). On the other hand, when a probability is clearly epistemic, in the larger scheme of things and not just in the fine details, we may call it an *evidential probability*. But always keep in mind that *(1)* an epistemic chance is *really* an epistemic probability, and not a true chance; and *(2)* we could legitimately refer to *any* epistemic probability as an evidential probability, if pressed.

4.3.1.2 Causal versus Empirical

Interpretations of probability can also be roughly divided into *causal* versus *empirical* viewpoints.

1. *Causal interpretations* attempt to explain probabilities in terms of the underlying causes that generated the events the probabilities are about, in the first place.
2. *Statistical interpretations* avoid invoking causes, and consider a probability to be strictly a matter of the statistics of observed outcomes.

It might seem that causal interpretations must be inherently metaphysical (since they invoke metaphysical "causes") while statistical interpretations must be purely epistemological, since they avoid causes and invoke only observation. However, this is by no means clear. It may be perfectly coherent to support a statistical view of probability, while still believing that the resulting statistic is an objective feature of the system under study. And it may likewise be possible to maintain a causal view, while believing that the actual assigned value of the probability statistic will have to take into account one's state of knowledge.

There is no implication intended here that causalists do not employ statistics, nor that the statisticalists do not believe in cause. While there may be underlying causes for the observed statistics, the statisticalist does not see these causes as part of what "probability" refers to. And while a causalist uses statistics to calculate probabilities, she believes it necessary to invoke the causes of those statistics in order to make sense of them *as* "probability".

4.3.1.3 Single-cases versus Ensembles

Most of the controversy over objective versus subjective interpretations revolves less around how objective we want our probability theory to be, and more around how seriously we take the idea of single-case probabilities.

Clearly, we all apply probabilities to situations that at least *appear* to be single cases, no matter what our interpretational inclinations:

- "The chance of *this* atom decaying within one hour is 1%."
- "The epistemic chance of *this* unfair coin coming up heads is 25%."
- "The evidential probability that *this* murder was committed by *this* butler is 38%."
- "My credence that my spouse cheated on me at *this* office party is 70%."

And likewise, there are situations where it at least appears that the probability refers to multiple cases, and *not* a single situation:

- "The chance that an atom of such-and-such an isotope will decay within one hour is 1%."
- "The epistemic chance of any given one of these unfair coins coming up heads is 25%."
- "The evidential probability that string theory is true is 38%."
- "My credence that my spouse will cheat on me at any given office party is 70%."

(Note that I've categorized string theory as multiple-case, whereas the butler's guilt is listed as single-case. This is because the butler's guilt is about a single incident, whereas string theory—if it is true—is a universal law that applies everywhere and everywhen in the universe.)

It would seem, then, that both kinds of probabilities are valid. However, when we look at the actual history of various probability interpretations, we will see that much of the debate hinges on the issue of whether single-case (or multiple-case) probabilities are even possible.

Note that when we go from the singular case to the plural, what was merely an *epistemic* chance (not a true chance) may, in fact, become a true chance. For instance,

- "The probability of *this* person having a heart attack in the next ten years is 0.1%."

is almost certainly an epistemic probability, since it is highly unlikely that we could know everything about the world that was relevant to whether this person would have a heart attack. However, assume that we know absolutely *nothing* relevant about this person (other than that they are a currently living person). This is still an epistemic probability. However, for certain facts about the world of this kind, when we pluralize the probability statement, it becomes a statement about objective chance, *not* epistemic probability:

- "The probability of any given person having a heart attack in the next ten years is 0.1%"

Even given that our state of knowledge remains exactly what it was for the singular statement, it is possible that this new plural statement might no longer be epistemic, but might now be objectively true of the world (presuming that we can make the terms precise enough, so there is no ambiguity as to who is alive in the world at any given moment, or what a heart attack is). At any given time, there is some number N people alive in the world, and some number h of them will have heart attacks within the next ten years (assuming no quantum effects, but even if there are quantum effects, presuming these are also objective chances, taking them into account will not make our probability any less objective). Thus the *objective* heart attack risk for a random pick from all living persons is exactly the same as the *epistemic* risk for a particular person for whom we have no relevant information:

$$p(\text{heart attack}) = \frac{h}{N} \tag{4.7}$$

The odd thing here, is that one could well argue that these two statements, one epistemic (and hence partially subjective), the other objective, represent the exact same *fact* about the world. Yet how can the same fact about the world go from being subjective to objective, just by re-phrasing it?

The correct response, I believe, is that they are *not* really the same fact about the world, even though we calculate them using the same numbers and the same formula, and get the same numerical answer. A plural statement about chances is saying something different from a singular epistemic statement, even if their truth is *justified* by the same fact about the world. There is nothing terribly unusual about this. I may justify the statements "there are twelve people in this room" and "there are an even number of people in this room" by citing the exact same facts, yet they are still different statements about those facts. To see clearly that the two statements are different in content, not merely in form, note that the epistemic statement's truth may change as we collect more information about the world, while the ontic statement remains true (or false) independently of however much I may know. For if I learn that the particular person referred to in the singular (epistemic) statement is a sixty-two year old man with a family history of heart disease, this may completely change my epistemic probability, but it will have no effect on the ontic probability. The two statements only

seem superficially the same because they are justified by the same fact *given* my current state of knowledge.

4.3.2 The Classical Interpretation

The oldest interpretation (that will matter to us) is the *classical* interpretation due to Leibniz [124] and LaPlace [64]. Classicists interpret the outcomes in the sample space as different *possibilities* for the same event. So the members of event ε are the different ways that the event *could* happen. The probability for event ε is thus a ratio of "possibility counts": the number of ways event ε can happen over the total number of ways anything can happen:

$$p(\varepsilon) = \frac{|\varepsilon|}{|\Omega|} \tag{4.8}$$

where $|A|$ is the size (or number of elements) of set A.

The classical interpretation is thus the most straightforward way of thinking of probabilities, and is closest to how we all thought about probability in school when we first calculated the probability of picking a red marble from a marble bag, or rolling a double-6 with a pair of dice. Other interpretations have mostly arisen and evolved as responses to perceived problems with the classical interpretation, or with the interpretations that later displaced it. Some of the perceived problems, however, will be of minimal importance to us in this dissertation, if they arise at all. Since I am not looking here to develop a general interpretation of probability adequate for all applications, my approach will be to start with the basic classical viewpoint, and deal only with those problems that actually concern us.

Most of the perceived problems with the classical interpretation can be classified into three broad categories.

1. *Infinities*: numerous problems and paradoxes arise when infinite sets are permitted.
2. *Indifference*: how do we know which sample space to choose for a given application?
3. *Singular cases*: how do we assign a nontrivial probability measure to a *single* case?

4.3.2.1 The Problem of Infinities

Problems of the first type concern infinities. Paradoxes and problems arise when we deal with infinite sets. For instance, if there are an infinite number of possibilities (whether countable or uncountable) in our sample space, then the ratio of possibilities will be undefined:

$$p(\varepsilon) = \frac{|\varepsilon|}{\infty} \tag{4.9}$$

While problems of infinities will be of some concern to us, they will not be the central concern that they are for some others, since we are starting with discrete computationalist assumptions, and our use of infinity will thus be highly constrained, and will not cause us the same trouble that it does for those who assume the reality of completed infinities, such as the continuum. However, even discrete probabilities involve infinite ensembles when defined classically, so we will still need to deal with the issue.

Nonparadoxical countable infinities One way of dealing with the problem (the most relevant way, for our purposes, that does not go fundamentally beyond classicism) is to concede that the classical ratio only applies to discrete cases, and to explain infinite cases not as actual probabilities, but simply as limiting cases. Let's take a very simple example that would seem to require an infinite sample space: the probability that a random natural number is divisible by 3. If $D3 = \{k : k \in \mathbb{N}, k \text{ divisible by } 3\}$, then we have the probability

$$p(D3) = \frac{|D3|}{|\mathbb{N}|} \tag{4.10}$$

Unfortunately, this is undefined, since the numerator and denominator are infinite. Yet, surely there is some sense in saying that the probability of a random natural number being divisible by 3 is $1/3$, since intuitively, it certainly seems that one-third of all natural numbers are divisible by three. However, this notion of a "probability" is not covered by discrete, classical probability.

Of course, we can get around this by defining such probabilities in terms of limits. We define the probability of a number in the first n natural numbers being divisible by 3 as $p_n(D3)$. This probability is *not* always $1/3$, but varies as n increases. Counting 0 as a natural number:

$$p_1(D3), p_2(D3), \ldots = 1, 1/2, 1/3, 1/2, 2/5, 1/3, 3/7, 3/8, \ldots \tag{4.11}$$

Thus, in the limit:

$$p_\infty(D3) = \lim_{n\to\infty} p_n(D3) = \lim_{n\to\infty} \frac{1 + \left\lfloor \frac{n-1}{3} \right\rfloor}{n+1} = \frac{1}{3} \tag{4.12}$$

giving the expected result in the limit.

Paradoxical countable infinities We could still argue, however, over whether $p_\infty(D3)$ should really be thought of as a probability *per se*, or whether it is best just to say it is the *limiting case* of a probability. After all, does it really make sense to speak of randomly picking a natural number? Just because the limit of a probability sequence exists, does not mean that the limit *is* a probability.

In favour of *not* calling it a probability is the fact that not all sensible finite probabilities can be so straightforwardly scaled up to infinite sample spaces. Take for instance, the probability $p_n(3)$

that a random number between 0 and n is 3, for some natural number n:

$$p_n(3) = \frac{|\{3\}|}{n+1} = \begin{cases} 0 & \text{if } n < 3 \\ \frac{1}{n+1} & \text{otherwise} \end{cases} \tag{4.13}$$

Now what is the probability that a randomly chosen natural number is 3? Let us again take the limiting case:

$$p_\infty(3) = \lim_{n\to\infty} \frac{1}{n+1} = 0 \tag{4.14}$$

Now if the limiting case of a probability is also a probability (so long as the limit exists) then this means that a randomly chosen natural number has *zero* chance of being a 3. But surely if this is true of 3, it is true of *all* natural numbers, including the one that we actually *do* get when we pick a random number:

$$\forall k : p_n(k) = 0 \tag{4.15}$$

which seems to be an obvious contradiction, and—in my opinion—as clear an indication as you could ask for that the idea of picking a random natural number is nonsensical. However, many feel perfectly comfortable with this apparent contradiction. They simply declare there to be a distinction between "impossible" and "zero probability".

"Just because something has zero probability," they will say, "does not mean it can't happen!"

The philosophical problem with this approach should be obvious. How can "zero probability" mean anything at all, if it doesn't exclude the possibility of the event's happening? The reality is, that it is only possible to make sense of such statements, ultimately, as limit expressions for probabilities. Such seeming contradictions can be avoided by (at least, in general) keeping the infinite limit of a probability as just that: a *limit* of a probability, and *not* an actual probability.

Since the notion of picking a random natural number does not seem to work for this case, the conservative approach would be to also reject it for the previous example of divisibility by three, as well, even though we got a sensible answer there. However, this would be (in my opinion) excessively pedantic. I will adopt the more practical convention of allowing the infinite limit of probability to be used as an actual probability, so long as interpreting it thus does not conflict with the more technically correct interpretation as a limiting case.

Uncountable infinities and the continuum The same kind of paradox arises in the case of the continuum, and other uncountable infinities. For instance, let's say we determine that the probability for a particle to be in the kth subdivision for a length that is divided into N subdivisions, is given by the function

$$p_N(k) = \frac{1}{N} \tag{4.16}$$

Define $\kappa(x, N)$ such that it maps a continuous position x onto a discrete subdivision, for a given N. Then define a continuous "probability", by taking this discrete probability to the limit, as N goes to infinity:

$$p_\infty(x) = \lim_{N\to\infty} p_N(\kappa(x, N)) = \lim_{N\to\infty} \frac{1}{N} = 0 \tag{4.17}$$

This gives us much the same paradox as the previous example. In one way, of course, it might seem to make some sense that if we placed a pin (with a perfectly fine point) down at an infinitely precise random location on a line segment, that the probability of hitting any one specific, continuous location *would* be zero. This is because, no matter where we place the pin, it seems that it must surely disagree with the pre-specified point at *some* level of accuracy, given that we can take the comparison to any arbitrary degree of precision we like. However, as in the last example, we are stuck with the problem that *all* continuous locations must likewise have had a prior probability of zero, including the one we actually *do* pick. We have to conclude, again, that the infinitistic value $p_\infty(x)$ should simply be considered a limiting case of a probability, *not* itself a probability.

More can be done to make such limits workable as probabilities *per se*. For instance, a "probability density function" is a real-numbered function that can be integrated over a range of continuous values to produce a nontrivial and nonparadoxical probability for that range. Such a function does not itself return a true probability (since there will be an infinite number of continuous values in any given range, and they would have to sum to unity to be probabilities). The function therefore needs to be integrated over the range to yield a true probability (the equivalent discrete function is called the "probability mass function" and *does* yield true probabilities for individual parameter values). However, it must be noted that, since the probability density function does not return a true probability for any *individual* real number, it cannot be used *prima facie* to support an infinitistic interpretation of probability, wherein it makes sense to talk about the probability of something being drawn from an uncountable infinity. No doubt, infinitists will find use of the probability density function compelling, perhaps even mandatory, but its use as a tool is in no way denied to those with a finitistic philosophy, since integration is not an inherently infinitistic concept, being ultimately defined mathematically as a *limit* on discrete computations. Since I will generally be sticking with a discrete philosophy in this dissertation, however, such devices will not concern us very much.

4.3.2.2 The Problem of Indifference

The Principle Any choice of a sample space in classical probability involves an assumption of indifference between the elements of that space. This is the Leibnizian/Laplacian *principle of indifference* [124, 64], a fundamental starting point for all applications of classical probability theory. It tells us that, if we have no reason for considering two or more outcomes in Ω to be any differently

probable, then we assume that they have the same probability. This stance could be due to metaphysical considerations (*i.e.,* the members of the sample space are the fundamental ontic entities of our system), or it could be a purely epistemic matter (*i.e.,* we simply do not know enough about the system to tell whether the members should have differing probabilities).

Principle 4.4. *The Principle of Indifference. Given n exhaustive, mutually exclusive, and indistinguishable outcomes $\omega_k \in \Omega$, then the probability $p(\omega_k)$ of the k-th possibility should be taken as*

$$p(\omega_k) = \frac{1}{n} \tag{4.18}$$

This principle tells us that these n outcomes can be treated as elementary entities, that can be counted equally in a probability calculation. In other words, the indifference principle has told us *what entities to count.*

The Problem While the principle of indifference has often been taken for granted, it has problems if treated as a single, straightforward rule. It is thus better to think of it as merely a starting point for the choice of a sample space, as it does not give us an unambiguous rule for choosing one. However, it does more than merely tell us something about an already determined space Ω—it *guides* us in our choice of Ω, in the first place. If our choice of Ω does not allow us to rationally accept a principle of indifference towards its outcomes, then it is not a very useful outcome set, since we cannot usefully count outcomes in order to compute probabilities, and this tells us that we are likely focussing on the wrong kind of thing as an "outcome". As a result, the principle of indifference is better thought of as a template with which to construct one's own specialized version of the principle, tailored to a particular application, rather than as a self-contained principle in its own right. From this point of view, we might rephrase the Principle:

Principle 4.5. *The General Principle of Indifference*

In constructing a probability measure, the sample space Ω should meet the following criteria:

1. Any two outcomes in the space $\omega_i, \omega_j \in \Omega$ are contradictory (mutually exclusive of each other).
2. The space is exhaustive (it includes all possibilities).
3. Any two outcomes in the space $\omega_i, \omega_j \in \Omega$ are, to the best of our knowledge, equally probable:

$$p(\omega_i) = p(\omega_j) = \frac{1}{n} \tag{4.19}$$

But this still does not lead us conclusively, in all situations, to the entities we should count. For there can be more than one way of defining the same set of possibilities in terms of outcomes, in which case even the generalized principle has a problem.

A standard textbook example [112, 20] used to illustrate this problem involves a bag in which we happen to know that there are exactly *three* marbles. All we know is that *(1)* there are three of them, and *(2)* that they are each either red or blue.

> ***Question***: *What is the probability that there are two red marbles and one blue marble in the bag?*

Some may find it surprising, but probability theory itself does not tell us the answer, for *it does not tell us what to count.* Even more surprising to some may be the fact that even invoking the general principle of indifference will not tell us what to count, in this scenario. Remember that probability theory itself is just a mathematical framework; it does not tell us how reality works—and, in particular for this example, it does not tell us how this particular bag of marbles came to be. For instance, we can define (at least) two different sample spaces for this problem, constructing both according to the general principle of indifference, and yet both will give different answers, depending on how the bags were actually prepared (so that the *history* or *origin* of the bag, not just its current *state*, must be taken into account):

1. The sample space Ω contains all possible ***combinations*** of red and blue balls (for a total of three balls). The n possible combinations (using B for Blue, and R for Red) are:

$$\Omega = \{3\text{B}, 3\text{R}, 1\text{B}2\text{R}, 2\text{B}1\text{R}\}$$
$$n = 4 \tag{4.20}$$

   Since we have no reason to prefer one combination over another, our *Principle of Combinative Indifference* tells us that the probability of any given combination, $c_k \in \Omega$, is the same as any other:

$$p\left(c_k\right) = \frac{1}{n} = \frac{1}{4} \tag{4.21}$$

   Answer: the probability of one blue and two red marbles (1B2R) is

$$p(1\text{B}2\text{R}) = \frac{1}{4} \tag{4.22}$$

2. The sample space Ω contains all possible ***permutations*** of red and blue balls (for a total of three balls). The n possible permutations of R and B (using B for Blue, and R for Red) are:

$$\Omega = \{\text{BBB}, \text{BBR}, \text{BRB}, \text{BRR}, \text{RBB}, \text{RBR}, \text{RRB}, \text{RRR}\}$$
$$n = 2^3 = 8 \tag{4.23}$$

   Since we have no reason to prefer one permutation over another, our *Principle of Permutative Indifference* tells us that the probability of any given permutation, $r_k \in \Omega$, is the same as any other:

$$p\left(r_k\right) = \frac{1}{n} = \frac{1}{8} \tag{4.24}$$

   We see here that, unlike for combinations, there is not a one-to-one correspondence between outcomes and permutations (which are our events). Hence, we need to create disjoint subsets

of the sample space to correspond to the events of interest (the permutations)*:*

$$\begin{aligned} \varepsilon_{\boldsymbol{3B}} &= \{\boldsymbol{BBB}\} \\ \varepsilon_{\boldsymbol{3R}} &= \{\boldsymbol{RRR}\} \\ \varepsilon_{\boldsymbol{1B2R}} &= \{\boldsymbol{BRR}, \boldsymbol{RBR}, \boldsymbol{RRB}\} \\ \varepsilon_{\boldsymbol{2B1R}} &= \{\boldsymbol{RBB}, \boldsymbol{BRB}, \boldsymbol{BBR}\} \end{aligned} \tag{4.25}$$

Answer: the probability of one blue and two red marbles (1B2R) is

$$p\left(\text{1B2R}\right) = \sum_k p\left(\omega_k \in \varepsilon_{\text{1B2R}}\right) = 3\left(\frac{1}{8}\right) =$$

$$p\left(\text{1B2R}\right) = \frac{3}{8} \tag{4.26}$$

We have acted, in both cases, in accord with the general principle of indifference, and looked for a sample space with indistinguishable members, and took them to be equiprobable. But we got a different answer for both cases.

The most sensible way to respond to the problem of indifference, within classicism, is simply to consider it a part of our job, in applying probability theory, to find the appropriate symmetries, for a given situation, that will group equiprobable entities together. Finding symmetries is arguably what science *does*, in its theory formation stage, so this may not be an unreasonable expectation for a scientist. On the other hand, if an alternative interpretation of probability could be found that was as satisfactory as classicism, except that it did not require us to do this work, then it would clearly have the upper hand. We will see later that this is just what the frequentists thought they had accomplished.

4.3.2.3 The Problem of Single-cases

Finally, there is the problem of singular cases. The classical ratio is a ratio of counts, and both the numerator and denominator are measures of the sizes of sets, i.e., ensembles. How, then, do we account for single-case probabilities, such as the probability of heads on *this* particular coin flip?

The pluralist response The classical answer is that apparently-singular cases are really part of an ensemble of different cases (either actually or hypothetically). Such a "classical pluralist" might allow that single-cases are possible for degenerate or trivial situations, but would argue that such examples can only yield probabilities of 1 or 0. In this view, a probability is never a probability of *this experiment* yielding *this outcome*, but rather is always (for nontrivial cases) a probability of *this kind of experiment* yielding *this kind of outcome.*

To understand this perspective better, let's take the four examples of single cases that I presented earlier, and see how a pluralist might reveal the implicit ensemble underlying each.

> "The chance of *this* radioactive particle decaying within one minute is 1%." *becomes*
> "The chance that the term (in the superposition describing *this* radioactive particle) that is actualized, after a one minute lapse, will be one of the terms describing decay, is 1%."

The pluralist would say that if this were really a single-case probability, the true probability would have to be either 1 or 0, depending on whether the particle actually *does* or *does not* decay after a lapse of one minute. If there is only one case (hypothetical or actual) that the probability is about, there can be only one outcome, so there cannot be an intermediate probability like 1%. In the case of radioactive decay, we are talking about an ensemble of possible outcomes, given by the quantum superposition. This is a legitimate ensemble to the pluralist even if he or she does not believe in the Everettian reality of all branches, since no matter your view of the reality of the multiple possibilities described in the wavefunction, the idea of there being a nontrivial quantum probability only makes sense in terms of such an ensemble of possibilities (whether or not the other possibilities in the ensemble are real or just abstract theoretical constructs).

> "The epistemic chance of *this* unfair coin coming up heads is 25%." *becomes*
> "The epistemic chance of *this* unfair coin, with an experimental setup of this type (described within our coarse-grained model of coin-flipping) will come up heads is 25%."

If we really were concerned with only one precisely defined experimental setup, the probability would, again, be either 0 or 1. We only get an intermediate probability because our coarse-grained model only describes the setup to a certain degree of accuracy. Therefore, all the possible (but clearly hypothetical) setups allowed by the coarse-grained model form the ensemble we are finding the probability of. There is no single case here.

> "The evidential probability that *this* murder was committed by *this* butler is 38%." *becomes*
> "The evidential probability that *this* murder, given a real-world situation of this kind (consistent with our available evidence) was committed by *this* butler is 38%."

Since we do not completely understand the real-world situation, we are actually dealing with a huge plethora of possible situations—anything at all that is consistent with our evidence. If we knew the situation exactly, we would surely already know whether the butler did it. Our ignorance results in an ensemble of possible situations.

> "My credence that my spouse cheated on me at *this* office party is 70%." *becomes*
> "My credence is 70%. that, out of all the possible things I think my spouse might have done at *this* office party, what they actually did included cheating on me."

Again, if there were no ensemble of possible scenarios, you would *know* whether you were cheated on.

I have belaboured the point perhaps more than necessary, in order to make absolutely clear that the basic argument of the pluralist applies to all of the four kinds of probability we discussed earlier (chances, epistemic chances, evidential probabilities, and credences). It is a general feature of probabilities (according to the pluralist's argument).

The singularist response Just as pluralists argue that non-trivial probabilities must be multi-case, the singularist will argue that non-trivial probabilities either *can* be singular [159, 142], or, even more radically, that they *must* be [8].

In a nutshell, the reason that single cases are still an issue of debate—in spite of the convincing sounding pluralist argument given above—stems from the fact that, due to the classical problems with infinities and indifference, there arose attempts to replace the classical interpretation with something better. The tension between causal and statistical approaches played a large role in this process, the end result of which was frequentism.

Unfortunately, frequentism ended up having more problems with single-cases than did classicism, which meant that we were returned back to having to account for single cases.

4.3.3 The Frequentist Interpretation

In frequentism, the classical ratio becomes a ratio of frequencies of actual experimental outcomes, in the limit of infinite experimental trials:

$$p(\varepsilon) = \lim_{n\to\infty} \frac{n_\varepsilon}{n} \tag{4.27}$$

where n is the number of experimental trials conducted thus far, and n_ε is the number of these trials with outcomes that are members of event ε.

While frequentism [77, 221, 222] would end up having trouble accounting for singular cases—and it is now widely felt that probabilities *can* apply to singular cases—its main original purpose was to avoid the problems of infinities and indifference, which were the most serious problems for the classical interpretation.

4.3.3.1 Indifference

Frequentism avoids the problem of indifference by remaining steadfastly empirical in nature. It does not address *why* a probability value will be approached in the limit, it simply defines probability as *being* this limit. This avoids metaphysical commitments, and the problem of finding some kind of symmetry on which to base indifference. However, this can also be seen as side-stepping the problem, rather than really addressing it. Surely, discovering symmetries is a core component of

theory formation, and a key part of the scientific process. Simpler theories are considered simpler largely *because* of symmetries (*i.e.,* principles of indifference). Thus, a classical derivation of a probability is surely based on a deeper understanding of the system under study than a merely empirical definition, *provided* we actually have a viable scientific model of the system that allows us to state a principle of indifference in the first place. But, then, surely it is only our ignorance that makes the frequentist definition attractive when we lack such a model.

A case can be made, then, that classical and frequentist probabilities are not really at odds with each other. We use the former when we have a sufficiently fine-grained model of the system we are studying, and we use the latter when we do not have such a model, but we *do* know how to describe the relevant experimental setup.

Whereas the classical definition requires us to isolate symmetries, frequentism requires only that we know how to set up the experiment in question, and how to recognize an outcome as an instance of ε. We can then find the probability $p(\varepsilon)$ simply by repeatedly performing the experiment over and over—except, of course, that to really find the *true* probability, we would need to do an infinite number of experiments, which is never actually possible. However, so long as we avoid the idea that the infinite sequence of experiments is supposed to represent an *actual* sequence—so long as it is a purely *hypothetical* sequence—then there is not necessarily a problem [222].

4.3.3.2 Infinities

Whether countably infinite or uncountably infinite, if our sample space or our events are of infinite size, then classicism has a problem (discussed earlier) as to how to make sense of the classical formula,

$$p(\varepsilon) = \frac{|\varepsilon|}{|\Omega|} \tag{4.28}$$

Early frequentists were unhappy with the answer we got stuck with in the classical interpretation—that we would just leave infinite cases as limits, and not define them as probabilities. The reasons they wished to save infinitistic probabilities is intimately tied into the infinitistic nature of classical physics, which drove the move towards frequentism, and which saw everything in terms of the continuum and real numbers. Frequentists therefore would like to believe that there really are objective probabilities (chances) that apply to infinite sample spaces.

Frequentism, instead of invoking limits only when infinities raise their ugly heads, embraces infinities and limits—as part and parcel of its definition of probability. Note the important differences over the classical conception. We no longer measure the size of a theoretically existing set. Instead, we simply perform experiments and note the results. Instead of defining criteria for set membership, we are defining criteria for recognizing the result of an experiment as an instance of event ε. We

are still counting something, but we are counting successful outcomes compared to total outcomes, rather than counting theoretical possibilities. We are no longer looking at the theoretical possibilities inherent in the situation. We are only looking at empirical results, not metaphysical causes.

The most obvious problem with the use of infinities in this scheme is the simple objection that one cannot actually perform an infinity of experiments. This problem was circumvented by von Mises [222] by requiring only a *hypothetical* sequence, not an actual sequence.

In summary, frequentism avoids the classical problems with infinities by:

1. incorporating limits directly into the definition of probability, and
2. sticking to a strictly empirical and statistical conception of probability (with the one proviso that the trials can be hypothetical).

When we applied limits in classical probability, we sometimes got paradoxical results, as when we asked whether a random natural number was equal to 3. Now, what if we viewed this same infinite sequence as a frequentist sequence of experiments? Suppose we perform a series of trials that has exactly this same sequence of successes and failures:

$$p_0(3), p_1(3), \ldots = no, no, no, yes, no, no, no, no, no, \ldots \tag{4.29}$$

The infinite limit then yields the same result we came up earlier for the classical interpretation:

$$p_n(3) = \lim_{n\to\infty} \frac{n_3}{n} = \lim_{n\to\infty} \frac{1}{n} = 0 \tag{4.30}$$

But here, such a limit is acceptable *as an actual probability*, where it wasn't necessarily so in the classical interpretation. Since a classical interpretation can feasibly be based on any model of abstract possibilities we can come up with, there is no guarantee that the infinite limit sequence will actually be a sequence of something that could ever really happen. The idea of randomly picking a natural number may or may not be a viable mathematical notion, but it is certainly not something one can do experimentally. However, if the very same limit can be interpreted, under frequentism, as a sequence of doable experiments, then we have, by frequentist lights, an appropriate infinitistic probability. Frequentism has gotten rid of the notion of an infinitistic random pick by tossing out the idea of a random pick altogether. We are no longer "picking" anything. We simply perform a specified experiment, over and over again, indefinitely.

4.3.3.3 Single cases

While frequentism may have improved on classicism in the areas of indifference and infinities, it falls short of classicism in the area of single cases. In the above example, we have a sequence of

almost all failures, except for *one* success on the fourth trial. So, while this does not generate the paradox we perceived under the classical interpretation, it does still seem puzzling. Since we have *defined* probability purely in empirical statistical terms, we must accept that there *could be* a single anomalous result like this in an experimental sequence. However, it somehow seems incorrect that a sequence that represents a probability could have such an anomalous result on a *single* trial. It would seem that the fourth trial must surely be a *different* experiment than the other trials. Thus, to make frequentism work, we are forced to define a set of *generating conditions* for our experimental setup and restrict our trials to those that meet these conditions [159].

However, if we must define conditions for our experimental setup, this compels us to peer into causes, which is not in the empirical-statistical spirit of frequentism, but is rather a causal approach, leading us right back to the problem of choosing a causal model, which is really just another variation on the problem of indifference. From the frequentist perspective, who is to say that one experiment "doesn't belong" in the sequence, since we are only permitted to look at the outcomes in the first place. To look at the experimental setups would be to distinguish between different causes, and we would no longer be taking a statistical, but rather a causal and non-frequentist approach.

And this brings us right back to the issue of single cases. One thing frequentism shares with the classical approach is the idea that there are no truly single-case probabilities, since there is always an (at least hypothetical) ensemble. However, in the sequence with the anomalous outcome, since the fourth outcome is completely unique, it stands to reason that the fourth trial *must* have a different *individual* probability than any of the other trials. It seems we are forced to accept single-case probabilities if frequentism is to work. Yet single-case probabilities can only work here if we allow ourselves to peer into causes.

4.3.4 The Propensity Interpretation

The propensity view of probability [159, 151] arose largely as an attempt to reconcile the frequentist interpretation with the apparent necessity of single-case probabilities. Popper [159] gives an actual example of an experimental sequence that, like the one in our example above, might produce a finite number of anomalous results. He asks us to imagine an infinite sequence of throws of a biased die that has a $1/4$ probability of rolling a six. In other words, the limiting frequency ratio is $1/4$. Now imagine that three *fair* throws are inserted somewhere into the sequence (we don't know where), each with a (single-case) probability of $1/6$ for rolling a six. The limiting frequency ratio for the *sequence* is still $1/4$, yet the *singular probability* for each of the three fair throws seems clearly to be exactly $1/6$. Thus are single-case probabilities thrust upon the frequentist.

Instead of trying to reconcile this problem with frequentism, Popper's solution is to reject fre-

quentism as untenable, declaring that any such requirements to delve into causes and models is no longer frequentism, and must be given a different name. Popper considers the generating conditions that define the nature of the repeated experiments in the sequence, and calls these causal forces "dispositions to produce limiting frequencies", or "propensities" for short. Thus was born the propensity interpretation of probability.

The interesting thing about propensities is that, even though they are defined in terms of limiting frequencies, and are simply a variation on frequentism (in one sense) they take us back to a view of things that is (in another sense) more like the classical interpretation, since probability is once again defined metaphysically, not purely empirically, as the strict frequentists would have it.

But then, is this really just the classical interpretation in disguise, where the "possibilities" used to calculate the classical ratio are now simply determined by the generating conditions? We could call these kinds of propensities "classical" or "possibilistic" propensities. But no, declares Popper, propensities *cannot* be based on classical possibilities, because possibilities do not come in varying degrees, which would be needed to get a propensity or nontrivial probability. There is nothing, according to Popper, about a "possibility" that could allow it to have a different weighting than other possibilities (that is, after all, why the principle of indifference declares all the possibilities to be equiprobable).

Of course, the classical answer to this is that, while *possibilities* are equiprobable, *events* (which are sets of possibilities) are not. The problem, however, is that, once again, we have something that fails to explain singular probabilities. Popper therefore rejects the classical approach to propensities. The problem here for Popper, however, is that singular events were only forced on him by *frequentism*, not by classicism, within which there is no such contradiction in the insistence on ensembles. So it seems that Popper ultimately rejects classicism and accepts propensities simply because he finds the idea of single-case probabilities compelling. The only really strong reason he puts forward for this, however, is the existence of quantum probabilities, which he takes to be strictly single-case. Indeed, he uses this as a basis on which to declare that his propensities can be considered new theoretical entities representing actual physical properties of things (akin, even, to physical force).

This is perhaps all rather extreme just to save the idea of single-case probabilities, when it is not even clear that we *need* to have single cases; perhaps probability simply always requires an ensemble of possibilities, at least hypothetical ones. While quantum probabilities may seem clearly single-case to Popper, this position is problematic (*especially* to Everettians!) in light of the pluralistic nature of the wavefunction. This presents a semantic problem, in fact, for this whole way of speaking: should we call a superposition of possibilities, before branching, a "single case"? Conventionally, this situation clearly *is* a single-case. But in the Everettian scheme, perhaps we should view it as already

multi-case, given the superposition. Clearly, we need to clarify the meaning of single and plural case probabilities, to avoid such confusion.

4.3.4.1 Strict and open singular cases

I will choose to reserve the phrase "single-case" to include Everettian interpretations of laboratory measurements that are conventionally considered single-case, for two reasons: *(1)* to do otherwise would seem to invalidate the whole idea of a "single-case", as the category would no longer even exist, and *(2)* counting different terms in a superposition as separate "cases" would seem to encourage the unjustified practice of branch-counting. Unless and until we accept some form of branch-counting, speaking about the pre-collapse wavefunction as anything but single case would seem presumptuous.

Of course, once branching has occurred, we seem to clearly have multiple cases, so we still need to qualify our use of "single-case".

Definition 4.6. A "singular" or "single-case" observation is one that requires one, and only one, conscious pre-observation observer.

Definition 4.7. A "strict" single-case observation is a singular observation that requires one, and only one, conscious post-observation observer (who is a continuer of—"remembers" being—the pre-observation observer).

Definition 4.8. An "open" single-case observation is a singular observation that requires an ensemble of conscious post-observation observers (who are all continuers of the pre-observation observer).

Definition 4.9. A "plural" or "multi-case" observation is any observation event that is non-singular—in other words, that requires an ensemble of conscious pre-observation observers.

Popper and the propensity theorists are clearly wed to the notion of strict single-case probabilities, which they are convinced are required, as Popper's arguments do not apply to the semantic grey area of open cases. We must always keep in mind, however, when considering propensities, that if some kind of pluralism could be saved, there would be no need for them. And Everettian observations provide this pluralistic element, in their open nature, even though we have chosen to continue to call them single-case.

Another problem for Popper's interpretation is that, while the idea of quantum probability—which Popper holds up as a standard for single-case objective probability—is based on a precise mathematical description of the physical system under study, the idea of propensity has no such foundation. Thus, Popper's propensities are, in general, some kind of mysterious physical what-nots about which one can say really nothing, except that they are *whatever it is* that produces the

limiting frequencies. Yet, if they are to be real theoretical entities, like forces and force fields, we need a mathematical explication of *how they generate* such frequencies. Theoretical entities have explanatory power; they are not just labels for observed frequencies, and so propensities fail as scientific properties of a system, and without this feature, they are little more than a re-phrasing of frequentism.

4.3.5 The Generative Interpretation

4.3.5.1 Accounting for single cases

Personally, I do not find the argument for *strict* single-case objective probabilities, in the Popperian sense, very convincing. Quantum probabilities are the only solidly single-case example of nontrivial objective probabilities that Popper seems to be able to muster, and they certainly *are* singular in a sense, but the wavefunction clearly presents us with an ensemble (of some kind, whether hypothetical or real), so I do not see why this should present a problem for classical pluralism (whether or not one is an Everettian). If the classicist wants to count up entities to compute a probability, open singular cases provide plenty to count (although exactly *what* to count will depend on our devising the correct principle of indifference). And, as we have already discussed, nontrivial objective probabilities for classical (non-quantum) events seem to *always* be pluralistic in nature. A single die toss or coin flip is not, for instance, really a single-case probability—except in the trivial sense of always have a probability of 1 or 0.

There are two possibilities for a given situation: either the system is (effectively) deterministic, or (for instance, if quantum effects matter) it is non-deterministic. If the system is deterministic, then a probability (for instance) of $1/6$ for rolling a six is not, in fact, an objective property of the system, and cannot be said to refer to some physical disposition or propensity. The true *objective* probability will either be 1 or 0, and our judgment of $1/6$ is merely a result of our ignorance of the full state of the system just before the toss, based as it is on a coarse-gained model (*not* a description of the exact state). But, as soon as we allow that our non-trivial (not merely 0 or 1) probability is a result of such coarse-graining, we have effectively introduced an ensemble of possible finer-grained models, via the coarse-graining. We may then define probability classically and pluralistically, counting the fine-grained models (or certain variables or entities generated by these models) as the possibilities in our classical probability ratio. We may still be faced with the issue of justifying our choice of what to count (the classical approach will always require this), but there is no need to consider such epistemic probabilities to ever be "single-case".

If the system is non-deterministic, then there might actually be an objective probability of $1/6$ for

rolling a six on a particular roll. Indeed, it is such objective quantum randomness that Popper used to bolster his insistence on having to account for single-case probabilities. However, from an Everettian perspective, these open cases are not strictly singular, any more than subjective probabilities are, since they arise from world-branching, which is the literal (but synthetic) realization of multiple possibilities. So, while it is possible that the idea of strict singularity for objective probabilities can be made sense of more generally (although I personally have my doubts), it can safely be ignored within an Everettian framework, in which there are always multiple realizations of any "disposition" of a quantum system. Thus, I will treat strict singularity similarly to completed infinities, and make the working assumption that they are not substantially relevant to quantum probabilities and the nature of the Born rule.

4.3.5.2 Generative Models

What I will call the "generative interpretation" has some things in common with Popper's propensity view, in that it requires a description of the generating conditions of the events in question. However, it does not use these conditions to generate an infinite experimental sequence. Thus, the generative view essentially invokes the "classical propensities" that Popper merely flirts with—he ultimately rejects them because he believes that singular possibilities cannot have different weightings. But if Everett is right, there really is no such thing as strict singularity, and the whole motivation for rejecting frequentism appears to be ill-founded.

However, Everett's "random pick" from an ensemble of possibilities (worlds) is only a random pick to begin with because Everett's supposed Born rule derivation claims to show that observers are the countable possibilities, and we have already seen that stage 2 of this proof is flawed. Thus, Everett's interpretation, as he left it, lacks any foundation for a generative or "classical possibilistic" interpretation.

Popper's objection to classical possibilism is ultimately just a restatement of the problem of indifference (as was the case for the frequentists, as well, with their objections to classicism). Popper claims there is no way to prefer one possibility over another, and I believe this objection can be addressed in a more general way than merely by invoking Everettian pluralism. I will describe another way around the problem of indifference for singular cases (a solution that is particularly suited to the approach I am taking in this dissertation). However, I stress that it is by no means the only possible solution. I will call it *generative probability theory*, as it is based on generative models (I make no claims here to its originality, as the use of generative models in probability and statistics is not new).

Clearly, the principle of indifference does not, in itself, tell us what to count. It is still up to

us to look at the real-world situation and decide how we want to model it. Different models will produce different measures. Returning to Keynes' marbles, recall that in the case of the combinative indifference principle, we are modelling the bag as containing a combination of marbles, and for the permutative principle, we are modelling it as containing a permutation. Which is the better choice? It may be tempting to choose the combinative principle, since it seems analytically simpler. The equiprobable entities are all of the same type as that in the question that was posed: they are combinations, and the question posed *was* about combinations. However, this is not valid reasoning. What matters is surely the reality of *how the marble bag was actually made—constructed, generated, set up or prepared*—not the nature of the *question* that is posed. We could just as well have been asked about the probability of a permutation, and the marble bag would still have been prepared in the same way. Moreover, we are apt to pose questions about our observations that align with our perceptions—in other words, with *appearances*. This is only natural, since perceptions are what we actually do observe. However, to assume that there is an isomorphism between these apparent entities and the true ontic entities that make up reality, is a fundamental error of reasoning: the confusion between appearance and reality. It would surely be highly anthropocentric to assume that the entities of reality must align with the way we happen to perceive them.

But if we are not to assume that the ontic countables of reality are the same as the apparent countables of perception, then we need a deeper model of the world than mere appearances, if we are to decide what to count.

Definition 4.10. A model of the underlying causes of something, where the entities in the model do not necessarily correspond readily to appearances, but which give rise to these appearances, is a *generative model* (since it *generates* the appearances).

Generative models tend to have a high degree of explanatory power, since they explain complex appearances as being generated from a simpler underlying model. They are to be distinguished from *discriminative models*, which only attempt to distinguish one thing, or class of thing, from another, not to model the underlying cause of those things.

Definition 4.11. A *pre-selection state* is the state of the system just before an outcome in the sample space is selected.

Definition 4.12. A *generator* is a model of the history of how the pre-selection state came to be.

Definition 4.13. If a generator is modelled with a step-by-step procedure for constructing the pre-selection state, then we call it a *constructor,* and the model is a *constructive model.*

Definition 4.14. A constructor that is modelled as an abstract computer program or Turing machine—or in some equivalent computational formalism—we will call an *algorithm*.

Depending on the situation, the generator may take the form of a laboratory preparation procedure, a cosmological theory, a step-by-step construction manual, or any of numerous other possibilities—whatever it was that made the situation what it was just before an outcome is selected. Later, I will argue that the kind of generator that is in general used in science is effectively the algorithm—but, for now, we will speak in more general terms about generators.

Returning to our marble bag, imagine there are two different ways the marble bag could have been generated/prepared, both using fair coin flips to decide between indistinguishables. Each generation method (or *generator*) will correspond to a different indifference principle:

1. *Combinative Generator*: the creator (generator) of the bag has two boxes, one with an infinite supply of red, and the other blue, marbles. She flips a fair coin *two* times. The first flip determines whether to choose marbles of all one colour (if heads), or a mix of colours (if tails). The second flip then determines whether the most frequent colour will be red (if heads) or blue (if tails). This generates the combinative Ω above.
2. *Permutative Generator*: the creator (generator) of the bag has two boxes, one with an infinite supply of red, and the other blue, marbles. She flips a fair coin *three* times, and places either a red (if heads) or blue (if tails) ball into the bag with each flip. This generates the permutative Ω above.

The permutative generator, and hence the permutative indifference principle, is the one that corresponds to how most of us would probably interpret marble-based probability questions in textbooks. But note that there is no *a priori* means to prefer it over the combinative interpretation, if the question asked does not give us the generative history of the bag, but only the partial information contained in a static description of the state of the bag right now.

Principle 4.15. ***The Generative/Constructive/Algorithmic Indifference Principles.*** *To apply the generalized indifference principle, create a generative/constructive/algorithmic model of how the pre-selection state came to be, and construct a sample space populated by indistinguishables from the model.*

There is no theorem, of course, that tells us to look for generators, nor proof for some other way around the indifference problem. However, like the principle of natural selection, the anthropic principle, or other guiding principles in science, it is an intuitively sensible and rational idea that guides our decisions. Not everyone may go along with us on the idea, but even those who do not, will see why the principle has intuitive appeal to some.

However, even this principle is inadequate, all on its own, to provide us with a probability rule. There may be more than one way to model the situation, for instance. So further assumptions will

have to be made, or empirical experiments done. Nonetheless, it seems that to get probability theory off the ground—particularly if we are to adopt an objective, discrete view of probability—then we need to make some assumption like this. In other words, if we want probabilities to be objective, we are going to have to *count* some kind of objective entity, so we will need to figure out *what* to count. The generative principle of indifference tells us to count the indistinguishables *in our generative model.* It does not, in itself, tell us how to construct the model.

4.3.6 The Bayesian Interpretation

Wishing to avoid the problem of indifference—so that probabilities might be "objective" and free from any prior metaphysical given—frequentists defined metaphysics out of the picture with a purely empirical definition of probability. But critics argued that this cannot be made to work for singular cases without a return to metaphysics. So the propensitists try to "fix" frequentism in a way that would work for singular cases, without reverting back to the pluralistic possibilism of the classical view. Unfortunately, this was unsatisfactory to many, who saw the definition of propensity as nothing more than an empty placeholder.

It would seem that if we are to go the metaphysical route, we need to find a *non-arbitrary* way to assign prior probabilities. And if we are to go the non-metaphysical route, we need to let go of the idea of defining probabilities in an objective way, since arguably "objectivity" will always lead us back to metaphysics. This thinking has lead the Bayesians to suggest that we embrace the arbitrariness of the prior probabilities—as doxastic probabilities, they are not objective measures, but simply degrees of subjective belief.

Doxastic probabilities, or credences, require no justification whatsoever from the state of the world or our knowledge of it (although they *may* be so justified). Often, we will say that credences need to be "rational" in the way we reason about them; and certainly, we must at least be constrained to reason about them in such a way as not to disobey the basic mathematical rules about probabilities, such as summation to unity and countable additivity (else we cannot call them probabilities). However, none of this means that all credences need to be actually justified by knowledge of the world.

One might counter this by arguing that a rational person *will* in fact justify his belief, so far as he possibly can, with evidence from the world, so that rational credences will never be *completely* unrelated to world knowledge, and hence probabilities will remain primarily epistemic.

However, the position that credences are necessarily epistemic in nature is difficult to make work. Recall that the mathematical framework of probability theory does not tell us how to assign all probability values, since there will always be some propositions that will need to be assigned

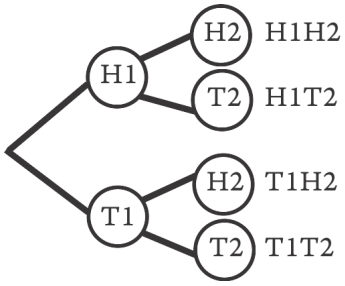
a prior value. In the classical interpretation, this is taken care of by the principle of indifference, but there is no necessity that dictates that we need to adopt indifference, especially since, as we have already discussed, we may not have any basis on which to decide what entities we should be indifferent towards.

The Bayesian approach assumes that some things are believed with a degree of certainty that is purely subjective. Even the principle of indifference doesn't really tell us why we should assign equal probabilities to the particular sample set that we do. So, why insist on indifference at all? Why not start with one's own initial subjective credence, even if it is just a gut feeling? One still must *reason* from these initial background beliefs in a rational way, and in a way that respects the rules of probability and logical inference. And, of course, if one really feels that one has no reason to prefer one choice over another, then it is still perfectly fine to start with equal probabilities. The principle of indifference is not unavailable to a Bayesian; it is just not mandatory.

Bayesians make extensive use of Bayes' rule to make inferences that combine their subjective probabilities to infer new ones. The rule can be interpreted from a classical or frequentist perspective, just as well as from a Bayesian one, so in spite of its name, it is not exclusive to those who are ideological Bayesians, although it is of particular importance to them, and that is why I will introduce it in this section. Bayes' rule starts with the notion of "conditional probability", where $p(A|B)$ is the probability of A given that we already know that B is true:

$$p(A|B) = \frac{p(A \wedge B)}{p(B)} \tag{4.31}$$

This is a fundamental relationship in probability theory, regardless of your interpretation. In the classical interpretation, it can be justified in terms of the equiprobable members of the sample space. Let's say that I flip a coin twice. There are four possible resulting sequences of heads (H) and tails (T), producing the following probability tree of possible outcomes:

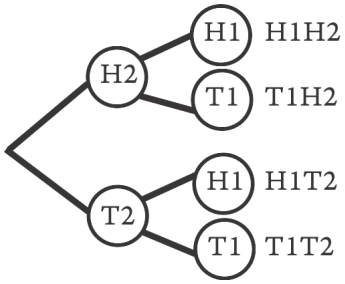


We see, for instance, that (assuming the appropriate principle of indifference), that $p(T1) = 1/2$,

and $p(T1 \wedge H2) = 1/4$. The conditional probability then follows from the tree:

$$\begin{aligned} p(H2|T1) &= \frac{p(T1 \wedge H2)}{p(T1)} \\ &= \frac{1/4}{1/2} \\ &= \frac{1}{2} \end{aligned} \tag{4.32}$$

We can also see from the tree that $p(T1 \wedge H2) = p(H2|T1)\, p(T1) = (1/2)\,(1/2) = 1/4$. Of course, since the two coin throws are independent events, the probability tree can just as well be written in inverse form:



giving us $p(H2 \wedge T1) = p(T1|H2)\, p(H2) = 1/4$.

which, combined with the definition of conditional probability, gives us Bayes' rule:

$$p(H2|T1) = \frac{p(T1|H2)p(H2)}{p(T1)} \tag{4.33}$$

When dealing with evidential probabilities, it can be convenient to think of this rule in terms of one outcome as an "hypothesis", H, and the other as "evidence", E:

$$p(H|E) = \frac{p(E|H)p(H)}{p(E)} \tag{4.34}$$

Logically, it doesn't really matter that we call one proposition the "hypothesis" and another the "evidence," but this is often how things are organized and conceived of, especially in science.

Bayes' rule is a rule of inference—a logical rule—that relates the probabilities of propositions H and E with the conditional probabilities of each proposition given that we already know the other is true. Let's say, for instance, that we wish to know the probability of having a heart attack within the next ten years, given a family history of heart disease:

$$p(\text{heart attack} \mid \text{family history}) = ? \tag{4.35}$$

We need to start with some "prior probability" for the hypothesis before we can attempt to take into account new evidence. Otherwise, we will not be able to get off the ground. Let's say we decide

that our credence for any given person's having a heart attack in the next ten years is

$$p(\text{heart attack})=1\% \tag{4.36}$$

Often, we do not know what the prior is, and from a Bayesian perspective, it is perfectly appropriate to simply give it a subjective value. If it happens that we *do* know what the prior is, then that's great, but the reality is that at some point, in our network of related propositions, there is going to be *something* we have no known prior for, and we will have to plug in some value or other, in order to make the inferential Bayesian machinery work. So we start with whatever our current degree of certainty of the hypothesis is, whether this degree of certainty originates from legitimate data, or is just a personal bias. If we look at how actual scientific inference works, this will always be true in the real world. New evidence that comes in for or against a scientific theory will always have to be evaluated against the backdrop of a person's given knowledge *and* their unjustified biases alike. But this fact does not prevent us from making sure that we reason *from* these priors in a logical and valid way.

Just as we need a priori probability for our hypothesis, we need one for the evidence, as well. Let's say that our credence for any given person's having a family history of heart disease is

$$p(\text{family history})= 20\% \tag{4.37}$$

Finally, we need to decide what we feel the prior probability of the evidence is, given our hypothesis. This is called the "likelihood" function, or "inverse probability". Let's say that we consider the probability of a family history of heart disease quite likely for a person who is actually going to have a heart attack:

$$p(\text{family history}|\text{heart attack})=60\% \tag{4.38}$$

We can now plug our prior knowledge/beliefs into the Bayesian machine to determine how the new evidence (of a family history) affects the hypothesis (ten year heart attack risk):

$$\begin{aligned} p(\text{heart attack}|\text{family history}) &= \frac{p(\text{family history}|\text{heart attack})p(\text{heart attack})}{p(\text{family history})} \\ &= \frac{(60\%)(1\%)}{(20\%)} \\ &= 3\% \end{aligned} \tag{4.39}$$

We have thus determined that the probability that a randomly chosen person will have a heart attack within the next ten years increases from 1% to 3% if we know that they have a family history of heart disease. The priors in the above derivation, according to the Bayesian interpretation of probability, do not need to be justified by appeal to empirical data. If I have no such data, then I set

my priors to whatever my subjective degree of certainty is. In fact, the above derivation is precisely an example of this, as I've no idea what the actual chances are of a random person having a heart attack in the next ten years. If I were to investigate and find out, I would be obliged to change my prior, but in the meantime, 3% is a valid derivation of what my rational credence should be, given my prior subjective beliefs on the subject.

The Bayesian has the problem of setting priors, just like the classicist has the problem of deciding on a principle of indifference. The classicist makes his choice based on reasons of theoretical simplicity. The Bayesian starts with what priors he has, whether they be due to a principle of indifference, previously collected data, or a vaguely understood personal conviction.

While many are still sceptical about the use of entirely subjective measures, much work has been done to show that this is a consistent approach, and that the subjectivity involved does not bring the system down; credences (at least rational credences) can and do behave like proper probabilities. In particular, credences have been defined in terms of betting and betting risk, using the cost-benefit definition of doxastic probability [167, 142]:

> To have a credence for A, CR(A) or $p_{doxastic}(A)$, means that $CR(A) = \frac{stake}{payout} = \frac{n}{n+m}$ is the highest cost to benefit ratio you would be willing to accept for a bet on A, where n is the stake or amount of your bet, and m is the amount put up by those you are betting against (equivalently, this means that $n : m$ are the worst odds on A you would be willing to accept).

"Dutch book" arguments can be used to show that such betting-based credences always lie between 0 and 1, and obey countable additivity and summation to unity, assuming the "coherence principle", which states (intuitively enough) that:

> *A rational agent will not accept a bet with a cost/benefit ratio that guarantees a loss.*

The betting-based definition of doxastic probability is not really restricted to betting situations, since the bet is merely hypothetical. To say I have 90% credence in string theory means that, *if* someone (God, perhaps?) could somehow settle the bet with certainty at some point, I would be willing to risk no more than 90¢ to bet on it, if this made for a total payout of $1.00 in the pot.

We are free to generalize the notion of a bet, so that the "cost" and "payout" are in non-monetary terms (such as "energy expended" and "food"), should we find the idea of defining probability in terms of money to be arbitrary. Ultimately, betting is just one example of the more general process of decision-making, since all decisions involve some cost and benefit to the decision maker. Hence, betting-based notions of probability are a common foundation for decision theory [186].

Of course, a "Bayesian", in the most extreme subjectivist sense of the word, doesn't merely believe that probability *can* be thought of doxastically, but rather that it *must* be thought of in these terms,

rather than ontically or epistemically. But if I can calculate the probability of plucking a marble out of a bag by counting marbles and doing a bit of long division, it would seem like overkill to talk in terms of credences, betting and subjective beliefs, when the actual objective situation seems so clear. However, Bayesians are, again, not barring the principle of indifference from use. They are rather saying that it does not provide a *general* platform for all useful talk about probability, and hence cannot be a foundation for our definition of what probability actually is. And, assuming we believe in (strict) single-case probabilities, there does not appear to be any appeal to objectivity on the table that would allow us to set prior probabilities in an objective way.

Quite aside from the response it provides to the perceived failures of classicism, frequentism and propensities, the Bayesian position has appeal on its own terms. It seems straightforward that, in at least some cases, we have nothing but intuition behind a subjective belief. It also seems in such cases, that we are still constrained to reason *from* this belief in a logical way, if we are to be rational. Moreover, Bayes' rule and betting theory give us reason to think this can be done in such a way as to be consistent with, and respect, the mathematical rules of probabilities—which means we would be hard pressed to say that such doxastic beliefs were *not* probabilities. Yet, if even *some* probabilities are entirely doxastic, *all* probability cannot be, by definition, objective; nor can it all be epistemic.

It could be argued, in favour of a pure subjectivist Bayesianism, that it is possible to view *all* probabilities as subclasses of doxastic probabilities, but not the other way around. Hence, epistemic probability is simply one type of doxastic probability, where the priors happen to be set by some justified model of the world. However, the Bayesian might point out that at some point, if we keep asking for more justification, we will find that even very well-justified epistemic probabilities rely on some kind of background beliefs that are not justified and simply doxastic. If our "objective" model was truly complete, the probabilities would all be 0 or 1.

There is one catch here, however, and that catch is quantum mechanics. In a deterministic universe, there should be no such thing as true single-case chances. But quantum mechanics seems to demonstrate the existence of objective chance for single cases. A rabbit trips on a twig, falls rolling down a hill, and bumps its head with a quantum probability of 80%. There is a 20% chance that the same rabbit will just barely avoid bumping his head, get up, and run happily off. What is the probability that this rabbit will experience the pain of the bump, as opposed to the relief of escaping uninjured? It seems difficult to deny that there is an objective quantum probability for the rabbit, even though the rabbit has no concept of probabilities, or the concepts required to even conceptualize of his situation in this way. Hence, the 80% probability is *objective*—independent of the beliefs of any observer, including the rabbit.

This would seem to be a problem for the Bayesian—at least those of the purest subjectivist

stripe, and certainly for those who would categorize quantum probabilities as subjective. On the other hand, to present the rabbit's probabilities as truly objective, one would certainly have to show a satisfactorily non-arbitrary way (at least in principle) of deciding on prior probabilities for apparently single cases, something the Bayesians—with considerable justification given the history of previous interpretations—tend to see as a fool's game. We could explicitly invoke quantum superpositions to justify open single-cases, thereby sneaking in a kind of limited pluralism. However, for now we are looking at the general interpretation of probability, so if there is a way to justify the idea of an objective single-case measure, without invoking quantum mechanics, this would be greatly preferred. And this is exactly what our next interpretation of probability theory attempts to do.

4.3.7 The Algorithmic Interpretation

4.3.7.1 Algorithmic Probability

The algorithmic interpretation of probability attempts to come up with a precise way of setting priors for *any* precisely describable *singular* structure, based on a classical "possibilistic" interpretation of probability, where probabilities are simply normalized counts of categorized ontic entities. It thus has ambitions as a means to setting objective priors in probability theory—although how "objective" these priors really are can be debated, and we will look at this question in more detail after we have come to a better technical understanding of what algorithmic probabilities actually are, mathematically.

The basic idea is to set the prior probability of a (possibly singular) structure or object according to its information content, which can be thought of as the number of bits it takes to communicate or describe the structure. Let's say I wish to describe to you something of great complexity: the sequence of bases in a particular DNA molecule, for instance. Let's say that you know enough about DNA that you know the alphabet of bases that such sequences are made of (usually abbreviated with the symbols G, C, T and A). However, you know nothing *a priori* about which kinds of sequences are more likely than others. Furthermore, if given partial information about a sequence, you have no way of knowing which completions of that sequence are more likely. In this situation, it would seem (by a principle of equivalence) that all possible sequences are equally likely, and should be assigned equal probability, since you are equally uncertain about each one.

Since our alphabet has 4 symbols in it, there are 4^n possible sequences, where n is the length of the sequence. The probability you assign to each sequence must therefore be $1/4^n$ or 4^{-n}. More generally,

$$p = b^{-n} \tag{4.40}$$

where b is the "base" of our code (the number of symbols in our alphabet).

In this situation, if I send you the first 3 symbols, I erase some of the uncertainty in your mind about which sequence you are receiving. We can say that you have received three "bits" of information about the sequence (although these are not the familiar base 2 "binary" bits, but base 4 "quaternary" bits, since $b = 4$). This means (assuming you are still completely ignorant about the remaining bits, and can infer nothing about them from what you have received thus far) that there are now 4^{n-3} different sequences you might receive. The probability of each remaining possible sequence is now $1/4^{n-3} = 4^{3-n}$, or in general

$$p = b^{r-n} \tag{4.41}$$

where r is the number of bits received.

Only b^{n-r} possibilities remain because we have eliminated b^r possibilities. This is equivalent to saying that $n-r$ bits of information remain to be transmitted, since we have already received r bits of information. Hence, there is a straightforward relation between information and probability:

$$H = -\log_b p \tag{4.42}$$

where H is the number of bits required to transmit a message with an *a priori* probability of p.

When you have received all the bits, then all of the uncertainty will have been erased. You will have received n bits of information, and there will only be one possible sequence remaining, whose probability will be exactly 1. The other previously possible messages are no longer possible, and now have probability zero.

This way of speaking about information is in terms of messages, information channels, and uncertainty. This is the traditional sense of information originated by Shannon [199]. Since it focusses on our uncertainty about the information source, in this way of speaking, "information content" is usually taken to be a property of the information source, as opposed to the individual messages, and represents an overall degree of certainty about which messages one might receive from this source. In the above example, the more messages that are eliminated as possibilities, the more certain we are about which message we will receive.

However, it is not necessary to eliminate any messages as possibilities, in order to receive information. One can receive bits of information without actually receiving any *particular* bits of the message sequence itself (at least not with any certainty). What if you did not receive *any* individual bits of the sequence, but were nonetheless given information that made it clear to you that certain sequences were much more likely than others? In that case, the total number of possible messages would remain the same. Yet the distribution of probabilities would no longer be flat (there would no

longer be a principle of equivalence for the set of all possible messages). There is still clearly a strong decrease in uncertainty. Just because none of the possibilities were literally reduced to zero, does not change the fact that their probabilities were decreased, and so overall certainty is still higher. The certainty, in fact, of the source is said to be the *average* number of bits it takes to communicate the message. This will be higher when the probability distribution is not flat.

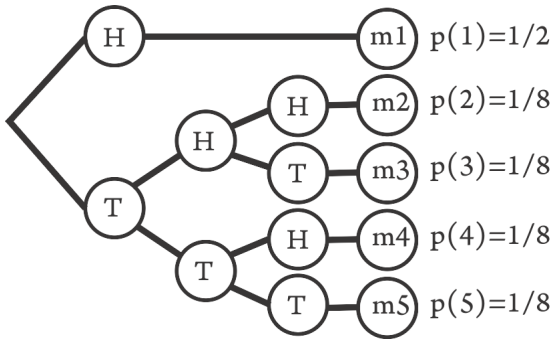
Algorithmic (or Solomonoff) probability takes a somewhat different approach to information theory, usually called "algorithmic information theory". It originated with Solomonoff [203, 204], in the context of a proposed solution to the problem of induction, and was further developed by Kolmogorov [119] and Chaitin [47, 48, 49, 50, 51]. The core mathematics of algorithmic information is formally very similar to that of Shannon information, with the biggest difference being a matter more of philosophical emphasis and interpretation. Algorithmic information focusses on the information contained in an *individual* object, rather than the source. This is not a hard-and-fast distinction, and there is no necessary conflict between the two formulations. However, with its emphasis on individual (singular) objects, algorithmic information is a more intuitive foundation for an interpretation of prior probabilities that happens to be largely concerned with single-case probabilities.

In algorithmic information theory, the information in a structure or object is the number of bits required to describe that object or structure. We can still conceptualize this as the number of bits required to erase uncertainty for the receiver. If we imagine it as the former, we are focussing on the idea that the individual object or message *itself* really contains these bits of information. If we focus on the latter view, we are imagining that the receiver of the message will have *his* uncertainty reduced on receiving the information. It would seem that the algorithmic view imagines that the object described has the information inherently within it, while the traditional Shannon view imagines that the issue is simply about transmitting the information to a receiver, and so the number of bits required will depend on the prior knowledge of the receiver. Shannon information, then, appears (at least superficially) to be based on epistemic probabilities, while algorithmic information appears (at least superficially) to be based on pure chance (ontic probabilities).

In reality, there is not so stark a contrast. Both information measures will need to factor in prior knowledge, when appropriate, and not even the algorithmic measure can get rid of the dependence, in an absolute sense, on prior knowledge. For instance, for Shannon information, if the receiver already knows that only two messages have any really significant probability, and considers those two to be equally probable, then it will take far fewer bits (on average) to transmit a message than if the messages were all considered equally likely. In fact, one could transmit the message with not much more than a single bit. By contrast, in algorithmic information theory, one focusses on the message as a singular object, and asks how many bits it takes, at minimum, to describe it. This

is not just the raw number of bits in a naive description. To find the actual minimum required number of bits, we need to consider *any* and *all* possible ways one might shorten the description. If a message is a sequence of ninety-nine 1's followed by a single 0, we could no doubt find a shorter description than listing all ninety-nine 1's. However, now we need to define our means of describing systems: our language. Some languages might be capable of more compactly sending some messages than others. So here we still have a dependence on prior knowledge, in spite of the different focus: there will be *some* kind of prior knowledge built into *any* descriptive language.

Both ways of looking at information are ultimately about efficient coding. To see this, let's consider a more detailed example. Imagine you are an information source, and each time you are polled, you transmit one out of a set of five possible messages: $\{m_1, m_2, m_3, m_4, m_5\}$. To determine which message to transmit, you flip a fair coin, and if you get Heads, you transmit m_1, if Tails, you flip the coin again. Then, if you get Heads, you transmit either m_2 or m_3 (determined by another fair coin toss) and if Tails, either m_4 or m_5 (also determined by a coin toss). Representing Heads with H and Tails with T, we can determine the probabilities, $p_1, \ldots, p_5$, for each of the five messages with the following probability tree:



What would be the optimal way to encode each of these messages into binary bits (0s and 1s)? You could, of course, encode them any way you wished. The first possibility that may come to mind would be to simply use the number of binary bits required to distinguish between 5 different objects, which would be 3 bits. Actually, 3 bits is more than we technically need, as it will actually allow for $2^3 = 8$ messages, and we only have 5. However, 2 bits will only allow for $2^2 = 4$ messages, and it would seem that we can't have a fraction of a bit of information (after all, how would we communicate "half a bit" of information?). However, in information theory, we often *do* speak of fractional bits. In part, this is because there are different units for measuring information, depending on the logarithmic base used. The "bit"—or more precisely the "binary bit"—measures information in base-2 logarithms, so we can distinguish between 8 different messages using $\log_2 8 = 3$ bits (and in general Ω messages with $\log_2 \Omega$ bits). But if we use a base-3 system, in which we have three

symbols to work with (such as 0, 1 and 2), then 8 objects now requires $\log_3 8 \approx 1.9$ "ternary bits" (or "3-ary bits") of information. Our 5 messages would require $\log_3 5 \approx 1.5$ ternary bits. Thus, sticking to our original choice of binary bits, the precise required number of bits for 5 messages would be $\log_2 5 \approx 2.3$ bits.

We are using base 2 as our default—which is standard in information theory—so when we speak of a "bit", without qualification, we mean a 2-ary or binary bit. Some other bases also have special names. A 10-ary bit is called a "ban", "decimal bit" or "dit". An e-ary bit is called a "natural bit" or "nat" (in physical applications, such as thermodynamics, nats are generally preferred over bits). From here on, I will go ahead and talk about fractional bits as if they were unproblematic, viewing them simply as an artifact of conversion between different units. If the idea of fractional bits still bothers you, just keep in mind that there will always be some base k that one could convert to that would measure the same amount of information in a whole number of k-ary bits (for our 5 messages, we would require a single 5-ary, or "quinary", bit).

Another way to view fractional bits would be that if you try to communicate 2.3 bits of information using 1s and 0s, and the receiver is allowed to guess which message you are sending after each bit, then you may get lucky sometimes and manage to communicate the message with only 2 bits, or even 1 bit. Other times, however, you may need 3 bits, but *on average* you will still require 2.3 bits to communicate the message.

So, does information theory tell us that you need 2.3 bits of information to properly communicate any one of your five messages? If the messages were chosen by the toss of a 5-sided die, then this would indeed be true, as all five messages would be equiprobable ($\forall k\, p_k = 1/\Omega = 1/5$). Here, we could re-use our logarithmic formula, using probability, p, instead of number of messages Ω. Since $p = 1/\Omega$:

$$\begin{aligned} H &= \log \Omega \\ H &= -\log p \end{aligned} \tag{4.43}$$

where H is the information contained in the message (or other physical system), Ω is the total number of possible messages (or system configurations), and p is the probability of any one of the (equiprobable) messages/configurations. (Note that the base of the logarithm may be omitted if it is immaterial or clear from the context.)

However, in the current example, all five messages are *not* equiprobable. Indeed, since you are transmitting m_1 fully half of the time, clearly it would be sensible to use fewer bits to send this message than all the others. Since you are transmitting the m_k message p_k of the time, a more efficient coding would be to simply traverse the above probability tree, with each 2-way branching

representing one binary bit of information (use 0 for H and 1 for T):

- m_1 : H = 0
- m_2 : THH = 100
- m_3 : THT = 101
- m_4 : TTH = 110
- m_5 : TTT = 111

Since we are taking advantage of the greater probability of m_1, this should amount to *fewer* than 2.3 bits per message, on average. Indeed, we find that the average message information, $\bar{H}$, is:

$$\begin{aligned}\bar{H} &= \frac{1}{2}H_1 + \frac{1}{8}H_2 + \frac{1}{8}H_3 + \frac{1}{8}H_4 + \frac{1}{8}H_5 \\ &= \frac{1}{2}(1) + \frac{1}{8}(3) + \frac{1}{8}(3) + \frac{1}{8}(3) + \frac{1}{8}(3) \\ &= 2 \text{ bits}\end{aligned} \tag{4.44}$$

Note that this is simply the average length in bits of each message, in the above encoding. It is not the mean of all the message lengths, of course; since the message lengths are not equiprobable, they must be weighted by the message probabilities. Generalizing, we have the following formula for the average information coming from an information source (or contained in a distribution of messages):

$$\bar{H} = \sum_k p_k H(k) \tag{4.45}$$

The information in a particular message is the same as when the messages were equiprobable, only now it differs with each message, since each message has its own probability p_k :

$$H(k) = -\log p_k \tag{4.46}$$

In other words

$$p_k = b^{-H(k)} \tag{4.47}$$

where b is the base of the (b-ary) coding scheme. This is called the "self-information" or "marginal entropy" of a message (or other physical system). Average information now becomes:

$$\bar{H} = -\sum_k p_k \log p_k \tag{4.48}$$

This is called the *average information* or *Shannon entropy* of an information source (or other ensemble/distribution of physical systems). Sometimes it is called the *information content* of the source, but I will avoid this language, as algorithmic information theory also talks about the information

contained in singular objects, and this language tends to confuse people over the difference between $\bar{H}$ (average over an ensemble) and H (information in an individual system or message).

Note that the lower the value of $\bar{H}$ (average value of H) the better the encoding scheme, since it will take fewer bits on average to encode a message. These values of H (and hence $\bar{H}$) are dependent, of course, on the encoding scheme. Our current scheme enjoys an average of 2 bits, which is certainly better than the 3 bits or 2.3 bits we considered previously. But is it optimal? Might there be a better encoding scheme? Examining the message codes, we see that no single message has 2 bits, but always either 1 or 3. Hence, we could safely drop the 1 from the beginning of all the 3-bit messages. This would certainly seem to be an optimal encoding for this distribution, giving us an entropy of only $\bar{H} = 1.5$ bits. This presents a number of minor nuisances, however; for instance, we now have one message (m_1) which is a "prefix" of two others (m_2 and m_3). This means when the receiver receives a 0 at the beginning of a message, he needs to know whether he will be getting another bit or not before he really knows that he has received m_1 or not. We could send some kind of "end of message" signal, but that will erase any gain achieved by dropping the leading 1s from the 3-bit messages. On the other hand, it is easy to see that as messages get longer, this end-of-message tag would become increasing insignificant, since it could be transmitted in a constant number of bits. Of course, we would also have to make sure that this special sequence (perhaps we will use "010") will be avoided in all messages, except at the very end. Some such scheme is also required if we want to use our code as a basis for sending larger messages, for which $\{m_1, \ldots, m_5\}$ will serve as a set of codewords, and larger messages can be sent, consisting of sequences of codewords. Without some means of indicating the end of a codeword, clearly, our message cannot be unambiguously (uniquely) decodable.

As it turns out, there are numerous matters of mathematical elegance favouring "prefix-free" codes. Not only does it mean we don't need a special end-of-message code (although such a code is one possible way, of course, to make the code prefix-free), but a theorem of Kraft [121] makes it much easier to deal with probability distributions based on prefix-free codes (also called "instantaneous codes" and "prefix codes"; note that, perhaps counter-intuitively, a prefix code is a prefix-*free* code, *not* a code with prefixes).

Theorem 4.16. *Kraft Inequality* [121][134, p76][179, p44]. *Let b be a natural number, and $H(1), H(2), \ldots$ be a finite or infinite sequence of natural numbers. Let H_{max} be the maximum number in the sequence. There is a prefix-free code with this sequence as lengths of its b-ary bit code words iff*

$$\sum_k b^{-H(k)} \leq 1 \tag{4.49}$$

Furthermore, if the code is "maximal" or "complete"—meaning that no more code-words can be added

of length $\leq H_{max}$ while retaining the code's prefix-free character—then the inequality becomes an equality:

$$\sum_k b^{-H(k)} = 1 \tag{4.50}$$

For complete prefix-free codes, the values in the above summation could feasibly take on the role of our message probabilities:

$$p_k = b^{-H(k)} \tag{4.51}$$

$$\sum_k p_k = 1 \tag{4.52}$$

Proof. (Sketch) An intuitive proof of the Kraft Inequality can be had by examination of the probability tree for the coding scheme, such as we drew above for our 5-message information source. Each branch in the tree represents one additional bit in a message, and thus each leaf represents a unique message. Wherever a message is shorter than the maximum (as with the "0" message in our example), the tree has been "pruned", so any further messages that might theoretically branch off from that message are unavailable (this is what makes it a prefix-free code). It is easy to see, by examining the probability tree, that if probabilities are assigned so that $p_k = b^{-H(k)}$, as they are in our example, and all available leaves are used in the code (the code is maximal), then the probabilities sum to 1, and the Kraft Inequality becomes an equality. If we exclude some of the leaves from our code, we are left with an inequality, as some of the probabilities in the distribution will be left out. However, the probabilities will still clearly sum to something less than 1, and the Kraft Inequality will hold in either case. □

Another advantage of prefix-free codes is that the Kraft Inequality can be shown to hold for uniquely decodable codes, as well as prefix-free codes [134, p77], which means that any uniquely decodable code can be replaced with a prefix-free code without changing the set of codeword lengths. Since any code must be uniquely decodable to be a satisfactory, unambiguous code, and any such code will correspond to some prefix-free code with the same $H(k)$ values, we can go ahead and restrict our consideration in algorithmic information theory to prefix-free codes.

Theorem 4.17. *Complete prefix-free codes are optimal iff $p_k = b^{-H(k)}$, where p_k is the probability of the k^{th} codeword* [134, p78]. *In other words, optimal codes will ensure that message lengths, $H(k)$, take on the values of the marginal entropies: $H(k) = -\log_b p_k$.*

Proof. (Sketch.) This should be evident from an examination of the probability tree. This is the coding that minimizes the expected codeword/message length, or Shannon entropy. In other words, *optimal codes are minimum entropy codes.* □

4.3.7.2 Analytic Recursion Theory

Algorithmic information theory focusses on the information contained in an individual object or message, so the probabilities in algorithmic information theory are more like the "single-case" probabilities that cause so much trouble in the interpretation of probability. However, in order to talk about the information contained in a singular object, one must choose a description language—or basis language—in which to describe the object. This basis language could feasibly contain a great deal of prior structure—"knowledge" if you will. This might make sense if we wanted to consider algorithmic probability to be a kind of *epistemic* probability. The prior knowledge of the observer is encoded in the language itself, and thus the information remaining in the "object" or "message" is the information, from the *subjective* perspective of that observer, given her prior knowledge.

However, our goal here is to develop a notion of *objective* probabilities for singular cases, so—while it is fine that we *can* take complex priori knowledge into account—in general we want to be able to assign rational *a priori* analytic probabilities to individual *analytic* structures. For this, we ideally want a language for general-purpose analysis that is not dependent on the particular knowledge that a particular observer or group of observers might have. The study of such languages is the subject of recursion theory. Recursion theory is the general theory of computational languages, which are the most general and expressive means we have of formally describing *anything* (anything, that is, which *is* formally describable). All quantitative scientific theories are ultimately analytic structures described by such languages—although, of course, the choice of which such structure applies to the real world may (or may not) only be knowable *a posteriori*.

I am ignoring the possibility that there are analytic structures that are *not* describable in computational languages. This is consistent with my assumption in this dissertation that computation is not fundamentally distinct from logic and mathematics; it is all analysis. However, to avoid confusion, I will use "analytic recursion theory" to refer to the study of recursive languages *as* models of analysis.

There are many different formulations of recursion or computation—many different computational languages—and we cannot just choose any such language. The whole point of developing an algorithmic notion of "prior probability" will be to define the prior probability in *a priori* analytic terms, so our language must be a *pure* language of analysis—or at least as pure as possible. It must thus satisfy, as closely as possible, two fundamental criteria.

An "analytic basis language" must have:

1. ***Maximal expressiveness***: be capable of describing any analytic object.
2. ***Minimal complexity***: be as free as possible (in syntax and in semantics) of synthetic artifice.

A language that does not satisfy criterion #1 will not be capable of describing some objects that are analytically describable, and so fails on the face of it to be a basis language for analysis. A language that fails criterion #2 will be too complex for our purposes. Since we want our language to represent analysis *qua* analysis, we cannot have any more synthetic artifact than is necessary (ideally, we would like to have none, but as we will soon discuss further, this will not be possible). Some languages may be maximally expressive, but too clunky and overly-complicated to truly serve as a universal analytic language. Others may clearly have some kind of specialized knowledge "built in". A language, for instance, that has the entire human genome available built-in ("for free," so to speak, as part of the language) is obviously not a *pure* analytic language, even if it is analytically maximally expressive.

It is possible to object that both of these criteria are impossible to meet for any real language (but this will not stop us from trying).

The Church-Turing Thesis Some may object that the first analytic criterion (maximal expressiveness) cannot ever be met, for how would we ever *know* that we had a language that could truly describe *anything* that can be precisely and logically described? Given any conceivable candidate for such a language, how would we know that there were not structures that could in principle be described, that our candidate language fails to cover? Perhaps there are whole new kinds of ideas that we *could* describe precisely and analytically, but we just haven't managed to think of them as yet. This is, in fact, a well-known issue in recursion theory. The working assumption that our standard computational languages *are* maximally expressive (at least with respect to what can be computed) is known as the "Church-Turing Thesis" [56, 217, 218, 220]. It is generally worded in terms of "computation" rather than "analysis", but since I am assuming here that these (along with mathematics and logic) are essentially the same thing, the Church-Turing Thesis (for us) is essentially the assumption that our standard computational languages are maximally expressive analytic languages (I will call this the "Analytic Church-Turing Thesis").

The thesis is sometimes spoken of (for instance by Turing himself [220]) as if it includes *all* computational processes (what we will call "programs"), including those that may not halt but continue on (theoretically) for an infinite number of computational steps. I will call this the "Analytic Church-Turing Thesis".

However, the thesis is more often spoken of as if it is restricted to processes that compute a required "result" or "output" in a finite number of steps. I will call this version of the thesis the "Church-Turing Computability Thesis".

However, since the class of maximally expressive languages for programs is identical to the class

of maximally expressive languages for computable functions (even though the latter is a proper subset of the former), the two versions are usually thought of as having mostly the same content. The computational Church-Turing thesis is always about programs, not computable functions; it is the most general version of the thesis, and the most relevant for our purposes, and so this is what I will mean by "Church-Turing Thesis", unless otherwise stated.

No version of the thesis is proven, and they are all quite likely unprovable. Nonetheless, the vast majority of those who work with computer languages on a daily basis believe very strongly in the truth of at least the computability version of the thesis (and I would dare guess that the vast majority of those who do not believe in completed infinities would accept the analytic version). This is not because they have any direct proof, but simply because after years of working with computer programming languages, one develops an intuitive sense that these languages capture everything we mean by "computation" (or "logically expressible" or "analytically describable").[38] There is a strong *a posteriori* component to this confidence, since part of why we tend to be so confident in the expressiveness of these languages is simply that they have been so widely used in information technology, to build so many practical and useful software solutions, used by billions of people, that it seems unlikely that we have somehow overlooked some class of structures that these languages cannot describe. I will generally assume, therefore, that the Analytic Church-Turing Thesis is true. Even if it is not, our current lack of solid reasons for believing otherwise makes a working assumption of its truth an eminently reasonable postulate.

The Simplicity Problem There are some who also may object that my second analytic criterion (for simplicity) cannot be meaningfully met by any actual language. All languages have some "built-in" structure, so how are we to say what part of that structure counts as "prior knowledge", and what counts as "pure analysis"? The answer is clearly that we cannot say—not absolutely. We are attempting to *define* analysis here, and we cannot use analysis to define analysis without a bootstrapping problem of some kind. Some kind of basic structure will have to be assumed as given, and our justification for it can only be based on its intuitive elegance and simplicity. We cannot measure this degree of elegance/simplicity—how would we do this other than analytically, and hence

[38] I am once again largely dismissing those who do not believe that computation is as powerful as analysis in general. The Church-Turing Thesis technically refers only to computation, so the analytic version is my own generalization, but one that follows from my assumptions in this dissertation. There are, however, many who see computation as only a subset of analysis—and in particular, many feel that infinitistic analyses escape the power of computation. While I disagree, I have generally avoided digressing into this debate in the text, and have assumed that "computation = analysis". However, for those who are interested in why I am so dismissive of the position that infinitistic mathematics goes beyond computation, I have included Appendices B and C, which define the technical concepts of recursion (Appendix B) and "limit recursion" (Appendix C). Appendix C, in particular, explains my reasons for believing that limit recursion renders computational languages fully expressive for the purposes of infinitistic mathematics, including all of number theoretic analysis and ZF (or ZFC) set theoretic analysis. A full treatment of the subject is beyond the scope of this dissertation. My own opinions on the subject are expressed more fully in [176].

by using the self-same basis language, the very thing we are trying to measure? For this reason, any attempt to measure the simplicity or complexity of our basis language seems to be doomed to circularity. Any definition of analysis, therefore, must necessarily be justified synthetically.

This sounds paradoxical, but it is not necessarily an unsatisfactory situation. In the philosophy of mathematics, it is generally accepted that any formal foundation must at some point rest on intuitive elegance. And if our ultimate goals are scientific, rather than mathematical (as they are, for us) the situation should be entirely uncontroversial, as some kind of intuitive criterion of simplicity is *always* necessary in science, to decide between theories (this is just Occam's razor, which is accepted as a basic postulate of science). Nonetheless, we still must do the work to show that our language of choice *does* meet some kind of reasonable standards for simplicity.[39]

Recursive Languages I will consider the class of maximally expressive analytic languages to be more or less equivalent to the abstract notion of a computer programming language. The "objects" or "structures" that we build with these languages—the entities we will be calculating algorithmic probabilities for—can be called "recursive structures" or "programs". The basic idea was first realized in the language of the Analytical Engine [143], the steam-powered device which would have been the world's first true computer if it had been completed. The idea was formulated in a more theoretical context in the "combinatory logic" of Shönfinkel [193] and [59], an attempt to provide a minimalist foundation for predicate logic (and probably still the most minimal of all analytic languages). Numerous other variations followed, including computable functions [93, 106], the "λ-calculus" [54, 55, 56, 57, 113, 114, 116], Post's formalism [160] and Turing machines [218, 217]. All of these ideas laid the foundation for the modern development of the digital computer and the many different computer programming languages that have been developed along with it.

These different languages were in time shown to be formally equivalent to each other. Computable functions are definable in the λ-calculus [116, pp 42-6], as are "Turing-computable functions" (those definable with Turing machines) [217, pp 230-40] . Computable functions have been shown to be definable in the λ-calculus [114], while any λ-definable function is definable with a Turing machine

[39]This comes with one proviso. I spoke earlier of the possibility that an Everettian quantum mechanics–especially if interpreted algorithmically–may encourage us to adopt a kind of idealism (a transcendental analytic idealism, to be precise). I am not fully addressing such a possibility in this dissertation (to do so would take us beyond quantum mechanics into general relativity and cosmology, theories not yet reconciled with quantum mechanics). However, I do discuss the possibility in several places, so it is worth pointing out that if we wished to defend a fully idealistic notion of quantum theory, we would not only have to unify it with general relativity, but we would no longer be able to dismiss the basis language problem in the name of Occam's razor. This is because a proper idealism would reduce all *a posteriori* knowledge to in-principle *a priori* knowledge, and Occam's razor is a rule of thumb for *a posteriori* science, not *a priori* analysis. This is not to say that simplicity could no longer be invoked, but if we wish to reduce science to analysis, the problem becomes equivalent to the problem of choosing an analytic foundation for mathematics or logic, and this is a more serious problem than the invocation of William of Occam in science, since science generally makes no pretension of absolute certainty, even in principle, whereas logic and mathematics generally do (at least historically they have).

[218, pp 160-1] and any Turing-computable function is, coming full circle, definable as a computable function [218, pp 161-3]. Therefore, we can use the phrases "Turing-computable function", "λ-definable function" and "computable function" interchangeably.

A computable function is one that actually has a defined value for any input arguments we give it; its computation *always* terminates or "halts" with an output or result. However, the languages used to define computable functions are the *same* languages used to define the more general class of "partial computable" functions (which includes those that may not halt and thus be undefined for some inputs). Thus the above result for computability straightforwardly applies more generally to partial computability (in other words, to computation in general). As a class, all these formally-equivalent languages are referred to as "recursive" or "Turing-complete".

Of course, in the years since these early results, many more languages have been shown to be Turing-complete. As a class, the Turing-complete languages represent our most comprehensive model of computation, and (I would claim) of analysis itself. While it is probably impossible to define an analytic language without *any* synthetic artifice, so long as we reduce this synthetic dressing to a minimum, we will have a strong candidate for an analytic basis language. One of the best ways to do this is actually to put forward several different candidates for our basis language, all maximally expressive, but conceptualized in very different ways. If we can understand what it is that these languages all have in common—whatever makes them inter-translatable and formally equivalent—we will have a far better idea of the notion of analysis *qua* analysis than we could ever glean from any one language.

Some important questions about recursive languages (at least from our perspective) are:

Question. *Is there is a more or less unique choice of Turing-complete language (or, more precisely, a class of languages that are trivially easy to translate between, making them intuitively identical) that is clearly preferred over (simpler than) all the others?*

Question. *Are the structures described by all the different Turing-complete languages objectively determinate (are they really the same objects)?*

Question. *Do the differences between Turing-complete languages result in the* same *probabilities?*

If the answer to any of these questions is a definitive "no", then algorithmic probabilities, while they may still be useful, cannot provide a measure of objective chances. It is thus important for us to not only choose an appropriate basis language, but also to show that its objects are determinate and produce unique probabilities. We will find shortly that there is a series of theorems from recursion theory and information theory that help considerably in answering these questions (although some of the answers are not entirely unambiguous).

I will start by listing eight different well-known Turing-complete languages, to give the flavour of the general idea of computation using different formal systems (but many more are possible):

1. *the Analytical Engine*
2. *first-order predicate calculus plus set theory*
3. *general-purpose programming languages (e.g.,* BASIC*,* C*,* Java*,* Lisp*, etc.)*
4. *partial computable functions*
5. *Turing machines*
6. *sequential Boolean logic (e.g.,* NAND *gates)*
7. *λ-calculus*
8. *combinatory logic (e.g.,* SK*-calculus)*

The list is roughly in order of increasing elegance (in my opinion). I will discuss here only the first and last languages on the list, but the interested reader can consult Appendix B, or a good textbook on computability or mathematical logic, such as [177, 134, 21], for more details on other languages. My own preference is for combinatory logic, the most common variant of which is the SK-calculus, which is remarkably simple and free of synthetic artifact.

It is extremely important to an analytic/computational epistemology that all these highly varied formal systems be inter-translatable, so that a solid conception of analysis can be developed that is not tied to any one specific choice of language. An expression in any of the above languages can be translated into an equivalent expression in all of the others. Hence, there seems to be a shared objective ontology for all these languages, consisting of just the content that survives this process of translating (between as many different examples of such languages as we can find).

The Analytical Engine While it is now mostly of historical interest, the language of the Analytical Engine must be mentioned, as it is the first analytically complete language, far and away pre-dating (at 1842) any other Turing-complete language in the literature [143]. Unlike the other languages in our list, the language of the Analytical Engine did not arise out of late nineteenth century developments in mathematical logic. It was, rather, the language designed for what was to have been the world's first general-purpose programmable computer, the Analytical Engine (invented by Charles Babbage), which was designed in detail and built in part, but never completed.[40] This steam-powered computer was to have operated in large part on the principles of the Jacquard Loom. As a result, it lacks the theoretical motivation and simplicity required for a basis language, being too complex and too obviously designed for a practical purpose.

[40] An online Analytical Engine emulator can be found at [224]. There is currently a project [96] underway to complete the engine, based on Babbage's extensive plans and using only nineteenth-century technology.

Its theoretical significance and analytical universality were, however, clearly recognized at the time. The chief programmer for the Engine was Lady Lovelace, who made it clear that the Engine was of use for processing, not just numbers, but any kind of symbols on which analysis can be performed, predicting future uses such as computer graphics and music generation. According to Lovelace, there is no pre-specified limit to the "powers of the Analytical Engine", which are "co-extensive with our knowledge of the laws of analysis itself." The Engine, in fact, is the very "material and mechanical representative of analysis," according to Lovelace. The language of the Engine she called a "new, a vast, and a powerful language... for the future use of analysis," a language which is not only the well-spring of all the "vast body of abstract and immutable truths" of mathematics, but also "the language through which alone we can adequately express the great facts of the natural world." The Analytical Engine was no mere numerical calculator, like its cousin the Difference Engine, which "is in its character exclusively synthetical," she wrote, but rather, "the Analytical Engine is equally capable of analysis or of synthesis."

Combinatory Logic Perhaps the most elegant formulation of recursion is combinatory logic [120, 193, 59, 61, 62]. It is also arguably the first fully recursive language without any obvious external semantics (with the possible exception of the Analytical Engine). It was originally intended as a kind of non-propositional, mechanistic foundation for the existing (propositional) system of predicate logic. A combinatory logic expression is a nested structure of ordered sets (lists) of symbols. Parentheses are normally used for the nesting. The primitive symbols of the system are operators that transform whatever expression they are applied to, usually appearing to the right of the operator.

The most common variant is the SK-calculus, which has exactly two operators: S and K. It helps to think of S as standing for "Synthesis" and the K for "projec-K-tion". SK-expressions are nested structures of these two primitive symbols, as in SKSKK or ((SK)SK((K)K)SK(S(SK)))—in other words, any sequence at all of S's, K's, and left and right parentheses, so long as the parentheses are balanced.

The S and K transformations are defined as follows:

$$Sxyz \longrightarrow xz(yz) \tag{4.53}$$

$$Kxy \longrightarrow x \tag{4.54}$$

where x, y and z represent any sequence of bracketed S's and K's with balanced parentheses. So, for instance, KSS is transformed into S in a single step, while $(S(KSK)(KK)(SS))$ evaluates to $S(SS)((KK)(SS))$ in two steps. Infinite recursion can occur here, as the evaluation may never terminate.

Given its lack of variable-bindings (as we find in most conventional logics), lack of any function-data distinction (S and K can act both as either function or as data) or any other obvious external semantics, and its extreme simplicity, I will take the SK-calculus as being very close to an ideal analytic basis language. (Within the context of ASU, this means that (4.53) and (4.54) constitute our current best model of ultimate reality; one could—literally!—not ask for anything simpler.)

What is Analysis?

Definition 4.18. A "computable" or "total computable" function is any (total) function that can be constructed out of SK-combinators. (A "total" function is one with a defined output for all possible inputs.)

Definition 4.19. A "partial computable" function is any partial function that can be constructed out of SK-combinators. (A "partial" function is one that may, or may not, have a defined output for any given input.)

Definition 4.20. A "recursive structure", "analytic structure" or "program" is any structure that can be constructed out of SK-combinators—or more precisely, whatever is invariant under translation of an SK-combinator into any arbitrary number of recursive languages (whatever is "recursively invariant"). An analytic structure can also be called an "analysis", and will be said to be an "analysis of" something when it is an abstraction of something synthetic (experiential). Whatever is part of the expression of a recursive structure that is not recursively invariant will be called the "synthetic artifact" of the expression.

Definition 4.21. A "computable algorithm" or "finite algorithm" is any program that can be interpreted as a computable function that solves a given problem.

Definition 4.22. A "computation" is any evaluation or "running" of an SK-combinator or program (or any process that is fully describable/definable by such an evaluation).

Note that the evaluation of a partial computable function is a computation, but that a computation may not halt, and hence may not actually compute a result. Hence, "computable functions" do not cover all possible "computations". However, the more general notion of "limit-computability" does cover all computation.

Definition 4.23. A function or problem is "1-limit-computable" if there is a program P_k that generates an infinite sequence of outputs, which converges on the required solution in finite time (whether or not it is computable *that* the output has so converged). A function or problem is "2-limit-computable" if it *would* fit the definition of 1-limit-computable if all the 1-computable functions

it uses could be replaced by computable functions. More generally, a function or problem is "n-limit-computable", for $n > 0$, if it would fit the definition of $(n-1)$-limit-computable if all the $(n-1)$-limit-computable functions it uses could be replaced with $(n-2)$-limit-computable functions (given that "0-limit-computable" is defined to mean "computable"). An algorithm is "limit-computable" if it is "n-limit-computable" for some n. A set is "n-computably enumerable" if there is an $(n+1)$-computable function for enumerating it. (See Appendix C for a more complete and formal definition of "limit-computable" and related concepts.)[41]

Just as the idea of "computable" needs to be generalized to "limit-computable" to cover all computations and recursive structures, the idea of "computable algorithm" can be generalized to "limit-algorithm" or just "algorithm".

Definition 4.24. An "algorithm" or "limit-algorithm" is any program that can be interpreted as a limit-computable function that solves a given problem.

I consider algorithms in general to be limit-algorithms, not necessarily computable algorithms, which is counter to standard usage[42]. Limit algorithms are not required to solve their problems in a finite number of steps (SK-combinator applications). They are permitted to solve them "in the limit" (see Appendix C for a detailed explanation of what it means to solve a problem in the limit). A simple example would be an algorithm that answers the question "Are there an infinite number of prime numbers?" by enumerating all natural numbers and testing each one for primality. It "solves" the problem if we are only concerned with the ontology of prime numbers. It just does not solve it in a finite number of steps, so it is limited as an epistemological tool.

Definition 4.25. A "computation" or "recursion" is any evaluation or "running" of a program (or any process that is fully describable/definable by such an evaluation).

---

[41] Terminology for limit computability is not as fixed as for computability. Some sources may define the default meaning of "limit-computable" to mean "1-computable", rather than n-computable, as I have done here. Putnam and Gold [165, 94], who originated the concept of limit computation, spoke this way, but their usage was restricted to computability and 1-limit-computability, so my usage is still a reasonable generalization of their terminology. Note, however, that the default for "computable" is always "0-computable" and never "n-computable".

[42] Standard usage is for the unqualified term "algorithm" to mean "computable algorithm" not limit-computable, as I have defined it here. I have departed from standard usage here for a reason. Conventional usage betrays conventional attitudes and biases, which are typically heavily biased towards terminating processes, even though these are the subset and not the general case. Hence, if we are concerned with recursive structures as an ontology, we *must* have an unconventional bias. There are historical reasons for the conventional usage, but we cannot be completely ruled by history. Recursion theory is greatly lacking in terms for computation where computation *in general* is the unqualified default sense of the term, but where we are nonetheless still computing a result or evaluating a function or solving a problem. If we are doing foundational work, this bias in the language can be mistaken for the actual state of affairs. I have tweaked the sense of "algorithm" because I believe it creates less clash with the literature than tweaking anything to do with functions. It is quite common for people to use the word "algorithm" more loosely than the functional terminology, anyway, and "algorithmic" is widely used informally as a near-synonym for "mechanical" or "realizable in a computer", which would be quite inappropriate if it did not cover nonterminating computations.

Definition 4.26. A "recursive language" or "programming language" is a Turing-complete language (any language or system in which all programs or recursive structures can, in principle, be constructed).

Once one has examined several examples of Turing-complete languages, and has learned how to translate between them, then one is in a position to describe what it is that they might have in common. I have only described the SK-combinators here, because I believe it is the closest thing we have to a pure analytic language. Hence, understanding this one language is a kind of short-cut to getting a good general understanding of analysis, but nothing really substitutes for understanding and learning several different systems, *and* a method for translating between them.

I will attempt to list some general features here that all analytic languages *must* have, but the reader should be mindful that this is a largely aesthetic exercise, influenced in large part by one's own preferences (in my case, I have a preference for combinatory logic). Thus, it is important to remember that *this list is not meant to be a formal definition of recursion*, nor is it meant to be a means for proving that any particular language is recursive. It is intended rather as an informal guide, summarizing my own experience and taste for what is elegant.

If you wish to rigorously demonstrate that a particular language is Turing complete, do *not* use this list, but stick with the traditional method: compile a complete translation manual between your language and some known Turing-complete language (see [21, Ch 6-8] for examples of this procedure).

With that proviso in mind, I will suggest that any Turing-complete language must have the following necessary and sufficient features: *structure*, *projection*, *synthesis* and *recursion*. Recursion theory takes its name after the final feature, but all four features must be present (the language need not be defined so as to explicitly acknowledge all of them in just the way I describe them, but the equivalent functionality and expressiveness must be there).

1. **Structure:** the language must have the ability to represent finite structures that are well-defined and unrestricted in size and the number of levels of structures within structures (within structures within structures...). This nesting of structures can be expressed in numerous ways, for instance, with punctuation (usually parentheses), or visually with graphs or tree diagrams. Structures are "transformed" into other structures by means of "operators".
2. **Projection:** the language must be able to break down structures into component structures, so it must have the equivalent of a "projection" operator that allows "truncating" (throwing away part of) a structure while keeping the rest.
3. **Synthesis:** the language must be able to synthesize or build up new structures out of component structures, so it must have some kind of synthesizing operator for combining structures. Following Shönfinkel [193], we will sometimes call the synthesizing operator the "fusion" operator (to distinguish it from synthesis in nonanalytic contexts). It must synthesize new

structures out of existing structures, while *at the same time* transforming them—true synthesis cannot consist of mere juxtaposition. The new structure must depend, not only on both the component structures, but also on some kind of interaction or relation between them. Metaphorically, one needs to "glue" the two structures together via a third structure (the "synthesizer" or "fusor" or "relation function") which is used *independently* to transform two structures that are then called on to act on each other, in order to generate the new synthesized structure.

4. **Recursion:** the language must be able to feed the result of a transformation back on itself (or into another operator that eventually feeds back into the original operator). There must be no limit to the number of times the application of an operator can thus recur (hence, there must be the possibility for some applications to recur forever).

In simple language, then, we could say that "analysis" is the recurrent process of building up and breaking down well-defined structures.

Criterion #1 describes the structures that an analytic language operates on. Criteria #2 (projection) and #3 (synthesis) describe the operators that perform transformations on these structures, while #4 simply requires that the operators be allowed to freely feed back on themselves without limitation. Note that I use the term "fusion" to mean "analytic synthesis", in other words, "synthesis of *well-defined* structures". Experiential structures are built up by the more general cognitive process of synthesis. Fusion is the building up of well-defined analytic structures in the human mind, which is only one particular kind of synthesis.

It may seem odd that a central feature of "analysis" is something I have called "synthesis", when analysis and synthesis are generally presented as opposing cognitive faculties. It may also seem odd that I have described the fusion (building-up) operator as "synthesis", but have not called the projection (breaking-down) operator "analysis"—even though in popular usage, "analysis" is usually said to be the "breaking down" of something, while synthesis is said to be the "building up" of something.

However, analysis clearly cannot *really* consist merely of breaking-down (projection) without any kind of synthesis. We commonly say that analysis is a "breaking down" process only because its most common application is the modelling of ill-defined experiences in well-defined logical structures. The ill-defined experiential structure is then said to have been "analyzed". Such a model has "broken down" the original non-analytic (ill-defined) *experience* into well-defined *logical* components. However, this "breaking down" cannot occur without using the experiential faculty of synthesis. One cannot "break down" something into logical components without *building* up those logical components into a model—it is all part of the same process. Indeed, logical structures can be constructed that are not intended to model or represent anything *else*, and it is more natural to think of this as largely a synthetic process of building up than as primarily a breaking down of something. But, of course, such building-up is equally necessary when one is representing something experiential.

Either way, analysis is largely—in practice—a process of synthesis, and I do not believe there is any reason to see this as paradoxical.[43]

There is no absolutely pure language of analysis, since analysis can only be realized synthetically. We will therefore never be absolutely certain that we have disentangled our analysis from the synthesis that was necessary in order to carry it out. Not that I think this is particularly a problem for the analytic method. Results of logical reasoning are no less certain, just because there is a fuzzy boundary between *(1)* what they mean, and *(2)* what was arbitrarily imagined in order to construct them. In Cartesian language, we would say that a logical outcome is "clear" (well-defined), but not necessarily "distinct", in that we can never say that *everything* about our expression of this result is really a necessary (as opposed to accidental) feature of it. Descartes saw rationalism as the commitment to make all our ideas clear *and* distinct [69]. Ordinary observations are both unclear and indistinct. Most mathematicians are very concerned with clarity but relatively apathetic towards distinctness (creating proofs that are exceptionally clear but very often spectacularly *in*distinct). My list of the four requirements for analyticity is an attempt to pay some service to the task of making our ideas *distinct*—and I believe distinctness is no less a requirement for rationality than is clarity—and modern recursion theory gives us reason to think that this is a reasonable goal, although one that can never be *absolutely* achieved (and we will see this fact demonstrated shortly, in some of the actual theorems of recursion theory).

If your favorite language does not readily meet the four analytic criteria, you may need to ask

[43] Of course, one might argue for the symmetrical case: that we should consider projection to be the general process of "analysis", which is also involved in synthesis as well, so that we can refer to the breaking down (projection) of purely experiential structures as "synthetic analysis". However, I don't find this to be a useful way to speak, as the symmetry that it implies between analysis and synthesis does not hold up. Analysis can only actually be performed by a rational being in an act of (experiential, non-analytic) synthesis. However, it is highly unlikely that the converse is true, that "synthesis can only actually be performed by a rational being in an act of (logical, non-synthetic) analysis". Perhaps a psychologist might successfully argue—and it may well be true—that, for any act of synthesis, its essential structure is the same as that of analysis, comprised of nested experiential structures, recursively acted on by projection and synthesis operations. This would mean that analysis was the *a priori form* of synthesis, but it would hardly be defensible to jump from this to the idea that we are actually ourselves performing an act of analysis whenever we synthesize. Of course, one might use the term "synthetic analysis" to mean something else entirely... as a way of simply stating that, in practice, all analysis is synthetic to some degree. And, while this is true, it would be confusing to thereby talk about "synthetic analysis". For one, the term would be redundant, since *all* analysis is realized synthetically (at least for humans). For another, it is misleading, since a cognitive process is an act of analysis only to the extent that one manages to extract its analytic form *from* its necessarily synthetic realization. Hence, while the synthesis involved in an act of analysis is no less synthetic because we are doing analysis, calling the act "analysis" implies that we are, at least to a great extent, able to generalize the act of synthesis out of its experiential context. There is not necessarily *any* analysis, on the other hand, in the act of synthesis. A bunny rabbit surely performs synthesis in every perceptual and cognitive act it performs, and yet it is arguable that it cannot perform analysis at all (and if you are thinking that its primitive reasoning capabilities perhaps do count as some kind of proto-analysis, it still seems clear that most of what it does is devoid of the analytic method; if you still doubt it, then replace the rabbit with the least rational animal that you still feel has the ability to synthesize its experiences from sense data). The necessary entanglement of analysis with synthesis is therefore not symmetrical. Analysis always involves synthesis, but synthesis need not involve analysis. This is the reason I use the phrase "analytic synthesis" for a basic cognitive function, but not the phrase "synthetic analysis", which I will reserve for the higher-level notion of analysing a system in synthetic terms. In other words, a reduction of a synthetic perception to purely analytic terms may rightly be called a "synthetic analysis" of that perception, but the analysis itself is no more or less "synthetic" in this case than in any other.

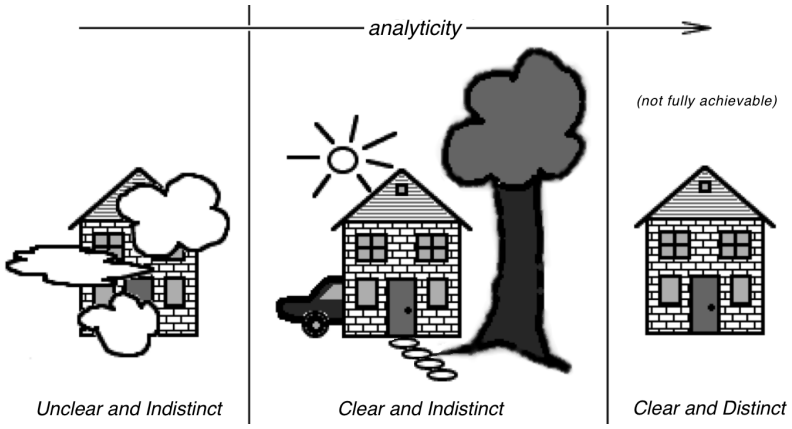


Figure 4.1: Degrees of clarity and distinctness of a house.

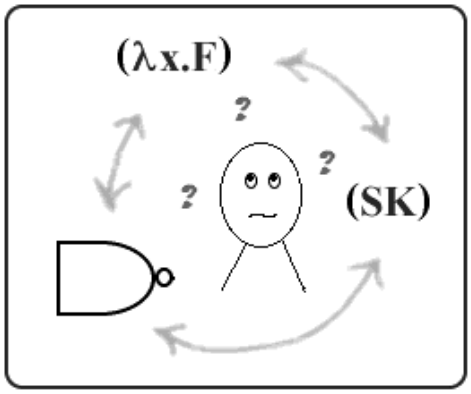


Figure 4.2: The universal language lies in the intersection of all languages.

yourself how these four criteria might nonetheless be possible to achieve within it. The functionality of the fusion operator is sometimes split across two or even more operators, for instance, and the functionality of the projection operator might be rolled into the fusion operator, making for only one actual operator.

The best way to see recursive structures for what they really are is not to rely on only one language, but to translate between the different Turing-complete languages, and see what drops out, and what remains. The truth to what computation, logic and mathematics are at bottom should lie in the intersection of these languages. Understand them all well enough to understand why a structure in one language is just another way of talking about the same structure as can be expressed in the others, and you begin to have a grasp for what analysis really is, *qua* analysis.

Assumption 4.27. *The Analytic Completeness of Recursion: the possible well-defined objects of human thought are exactly the recursive structures.*

Definition 4.28. By "possible" objects of human thought (or "ideas"), we include all those that are actually possible for a human to precisely think of, as well as all those that exist "in principle", *i.e.,* by implication from the existence of the actually thinkable ones. (*I.e.,* we are not excluding ideas that are not thinkable in practice, due solely to cognitive limitations of our finite human minds.)

The analytic completeness of recursive languages is a fundamental assumption of this dissertation, and if the assumption holds true, then all the possible well-defined objects of human thought are nothing more than recursive structures. Hence, we will assume further

Assumption 4.29. *The Assumption of Analytic Probabilities: Given the analytic completeness of recursion, the countables of an objective a priori probability theory (if there are any) will have to be the analytic (recursive) structures.*

The reason I am belabouring the point about choosing a *distinct* analytic language (one that encourages the expression of distinct recursive structures) is that it is typical in mathematics to focus only on clarity, and I do not believe we can adequately justify an algorithmic interpretation of probability theory without addressing distinctness. Clarity is all well and fine—if one wishes only to prove things and then apply the results—but distinctness is required if one wishes to truly understand such results on their own terms (in short, if one wants to do any kind of *foundational* study).

As long as a language is well-defined, the structures we construct with it will be clear. So long as the language is Turing-complete, we are confident that we can express all of them. However, if we pay no mind to their distinctness, we will fail to grasp that an expression in ZFC set theory is formally the *identical* structure to some expression or other in the SK-calculus. Since the synthetic artifacts of these expressions are so intuitively different, it may be irresistible to think that they *must* be truly different structures. This ultimately may lead to a rather fuzzily-understood belief in something "pointed to" by the recursive structure that is "outside of" the formalism—such quasi-mysticism (*e.g.*, [153, 137]) is almost inevitable when we ignore distinctness. Since there really *is* no formal difference between (say) a proof about a ZFC set and the application of the corresponding SK-expression, if we make no attempt to translate between the two languages, what else are we to do with the synthetic artifact but assign to it some kind of substantive "extra-formal" existence? Only translation between multiple Turing-complete languages will prevent us from falling into the quasi-mystical trap, and that is why I have mentioned a number of different languages here, rather than simply describing my own favorite.

None of this is to say that I place all these languages on an equal footing. I have already gone on record here that I think combinatory logic to be the most elegant formulation I have encountered.

However, this is not a dogma, but a provisional aesthetic judgment. For you, it might be something different. However, none of what I am about to develop, by way of an algorithmic theory of chance, will hinge on exactly *which* language you deem the simplest. It *will,* however, most certainly hinge on your ability to recognize that some languages *are* simpler than others.

My requirements for an analytic basis then, are, in short:

1. ***Turing-completeness***: the language must be capable of expressing any well-defined idea.
2. ***Simplicity***: the language must have minimal structure beyond the four analytic criteria, implying*:*
   Concreteness: there must be no built-in abstraction or variable-binding feature.
   Non-functionality: there must be no built-in "function-data", "input-output", "structure-operator" or other analogous distinctions.

The Problem of Determinacy Some may argue further, that there is no way I can even know that there *is* any coherent language at the intersection of all other languages. Were it possible to translate *any* structure in one language into *any other* structure in another—depending merely on which translation scheme we chose—then we could legitimately claim to have "deconstructed" recursion theory, and we would have reason to question its objectivity. However, no one has ever done anything like this. Those who work in this field are continually impressed, rather, by the constancy of certain aspects of recursion that are the same across many multiple languages. There is certainly redundancy, but no one has ever shown that this redundancy can be made to take over the entire class of languages. Certain classes of expression in one language map to analogous classes of expressions in other languages, and these we take to be our objective objects of human thought. While different translation manuals will produce different classes and hence different correspondences, it nonetheless seems to be the case, for example, that the structure that computes the 1000th prime number exists in all these languages, and the computation that is realized in each of them is essentially the same. This is, of course, an *a posteriori* argument, based on the experiences of many people; it is not an *a priori* proof. However, we will see shortly that, although we have no absolute *a priori* proof of determinacy of meaning, there are a series of theorems that *do* prove, *a priori*, some remarkable facts about recursion that, taken together, strongly imply that recursive languages are deconstruction-proof. These formal theorems will also illustrate many of the features of analyticity that I have thus far informally discussed (such as the entanglement of analysis with synthesis, and why it does not present a fundamental problem for the analytic method).

Kleene's Recursion Theorems

Kleene developed his own version of computable functions and showed them equivalent to Gödel's

earlier formulation and to Church's λ-definable functions and generalized them to the partial computable functions. Kleene also proved a series of theorems fundamental to recursion theory: the recursion and fixed-point recursion theorems [115], and the enumeration theorem [116].

Before we look at some of Kleene's theorems in more detail, we will define what it means for a set to be "computable" and "computably enumerable".

Definition 4.30. A set of natural numbers is said to be "computable" if there is a computable function (SK-combinator) that will tell us whether any arbitrary natural number is in the set or not.

Definition 4.31. A set of natural numbers is "computably enumerable" if there is a computable function that can tell us what the k-th number in the set is, for any natural number k (under some enumeration of the members of the set). In other words, the members of the set can be enumerated by an SK-combinator (for an infinite set, the combinator will be non-halting).

Theorem 4.32. *A set that is both computably enumerable, and whose complement (the set of non-members) is also computably enumerable, is computable.*

Since the set of all partial computable functions (or combinators or programs) is computably enumerable, we can index them. Let the following represent an enumeration of the 1-ary (one-argument) partial computable functions (or combinators or programs):[44]

$$P_1(x), P_2(x), P_3(x), \ldots, P_k(x), \ldots \tag{4.55}$$

Likewise, let the following represent an enumeration of the 2-ary (two-argument) partial computable functions:

$$P_1(x_1, x_2), P_2(x_1, x_2), P_3(x_1, x_2), \ldots, P_k(x_1, x_2), \ldots \tag{4.56}$$

The input arguments are all assumed to be in $\mathbb{N}$ (the natural numbers). Any pair of natural number arguments, (x_1, x_2), can be mapped onto a single argument $x =< x_1, x_2 >$, using an appropriate "Curry" function $< x_1, x_2 >: \mathbb{N}^2 \rightarrow \mathbb{N}$ (this is a procedure known as "Currying"). Thus, we can see that there is no real distinction between the partial computable functions of different numbers of input arguments. In fact, we can take $P_k(x_1, x_2)$ as shorthand for $P_k(< x_1, x_2 >)$. Likewise, we could define a 3-ary Curry function $< x_1, x_2, x_3 >$ to define $P_k(x_1, x_2, x_3) = P_k(< x_1, x_2, x_3 >)$.

[44] I use "$P()$" to represent programs, combinators, computable functions, *etc.* (think "P" for "program"). I sometimes use the term "computable function", since it is a very common term in recursion theory (traditionally "recursive function" was used, but "computable function" is gradually replacing it). However, the notion of a combinator or program is really more general than the notion of a function, so functional language can be confusing. In combinatory logic, an SK-combinator can be *interpreted* as a function, but can also be treated as data and passed as an argument to another "function". Hence, I prefer "program" as the most generic term ("combinator" might be even better, but is not as readily and widely understood).

So we can use this scheme to assume that all partial computable functions are 1-ary, and use "P_k" to refer to the "k-th partial computable function". We could also, however, use 0-ary partial computable functions, which would just be equivalent to the partial computable function calls or applications (the programs). Since the input arguments can be enumerated, so can the calls to the partial functions:

$$P_1, P_2, P_3, \ldots, P_k, \ldots \tag{4.57}$$

These can still be considered a kind of 1-ary function, however, since a 0-ary function can just be given an extra "dummy" argument that has no affect on its operation (for instance, define $f(x) = 3$). What all this effectively means is that it is a synthetic matter of interpretation how many "arguments" we consider a combinator to have, or even whether it is a function at all, as opposed to just a mechanical program.

Theorem 4.33. ***Kleene's recursion theorem*** *states that there is a computable function f, such that $\forall x \in \mathbb{N}$, if P_x is total, then*

$$P_{P_x(f(x))} = P_{f(x)} \tag{4.58}$$

In other words, there is a computable function f that will, for all computable functions P_x, provide (given x) an index to a program $P_{f(x)}$ that will exactly simulate the behaviour of program $P_{P_x(f(x))}$ indexed by passing the index $f(x)$ of this new program to P_x itself.

Thus, the gist of the recursion theorem is that algorithms can be simulated by other algorithms. An easier to understand and special case of the recursion theorem is

Theorem 4.34. ***The fixed-point recursion theorem****: for every computable function f, there is an $x \in \mathbb{N}$ such that:*

$$P_x = P_{f(x)} \tag{4.59}$$

In other words, for every computable function f there is an index (or "Gödel number"), x, of a program, P_x, such that the result of applying the function f to the index of P_x is the index of a program that exactly simulates P_x.

From the point of view of computer programs, f is sometimes thought of as a "code-editing" function, since it transforms one program P_x into a new program $P_{f(x)}$. The fixed-point recursion theorem tells us that for any such code-editor, f, there is always a program, P_x, that is unaffected by the editing. P_x is a "fixed point" of f.

Of course, there are f functions for which this is trivial and uninteresting. If we defined $f(x) = x$, we would trivially have a fixed point—for each program P_x, the fixed-point is the very same program P_x. This would say nothing more than that every program "simulates" itself (which is true, but

uninteresting). However, the theorem is about *all* code-editors (all total functions f). So no matter what strange, loopy and convoluted code-transformer we come up with, there is always some program that is its fixed-point—that it fails to really change.

The recursion and fixed-point theorems lead straightforwardly to:

Theorem 4.35. ***Kleene's enumeration theorem****: in any Turing-complete language, there is a program that enumerates all the programs.*

I will introduce this theorem in terms of computable sets. Let $B(m, n)$ be an arbitrary computable function that takes natural numbers m and n as inputs and produces another natural number as output. P_x is, as before, the x^{th} program. Let O_x be the set of numbers that, by some (arbitrary) interpretation of the machine's states, are produced during the running of P_x. The enumeration theorem states that the set of all values produced by $B(m, n)$ is computably enumerable if the first argument m is a member of the set of numbers produced by the second argument's combinator (the nth combinator):

$$K = \{B(m, n) : m \in O_n\} \text{ is computably enumerable.} \tag{4.60}$$

Members of set K are natural numbers, each of which represents a relation between two other numbers. This is just a way of stating something about the enumerability of a two-place predicate via a set of natural numbers. If $\{B(m, n) : P(m, n)\}$ is computably enumerable, then we say that the two-place predicate $P()$ is also computably enumerable. Thus, the enumeration theorem states that whether or not "m $\in$ O_n"—the question of whether a certain number is or is not ever produced by a certain combinator—is itself computably enumerable. Recall that this does *not* mean that it is computable whether or not $m \in O_n$. Computable enumerability means that we can search through all the numbers that Turing machine T_n *does* produce, looking for m, and printing out m if we find it. We can do this for all m and for all T_n. Every m that is produced by every T_n *will* get printed out, but if there is an m that is *not* produced by a certain T_n, we can never be sure, simply by enumerating O_n like this, whether m is in O_n or not.

A straightforward consequence of the recursion and enumeration theorems is that there is necessarily some redundancy in any Turing-complete language. In fact, one consequence of these theorems is that any Turing-complete language has a "universal simulator" or "universal Turing machine".

Definition 4.36. A "Universal Turing Machine" (UTM) is a program $T()$, such that,

$$T(k, x_1, \ldots, x_n) = P_k(x_1, \ldots, x_n) \tag{4.61}$$

Turing proved that such a function/machine exists [217, pp 241-6]. Note that if we assume there is no distinction between function and data (which we should be able to do in pure analysis), we can

rewrite this as:

$$T(k, x_1, \ldots, x_n) = k(x_1, \ldots, x_n) \tag{4.62}$$

In computer science terminology, this means that any programming language can be used to write a version of its own interpreter.

Isomorphism and Program Identity The inherent redundancy implied by Kleene's recursion theorems raises the problem of program identity: the concern that meaning in recursive languages might be indeterminate, making such languages "deconstructible" to the point that their objective content disappears, in spite of the inter-translatability between them. Rogers' isomorphism theorem [178] addresses this problem to a large extent, at least within the context of functional interpretations of recursive languages.

To understand the problem of program identity better, imagine that program X in language a maps to (*i.e.* can be translated to) program I in language b, according to some translation manual θ, which we will write as

$$X_a \longrightarrow_\theta I_b \tag{4.63}$$

and suppose the X_a, under some appropriate interpretation of its states, can be considered to be the partial computable function f_x. Put another way, choosing f_x arbitrarily fixes (assigns or "baptizes" certain states as representing) the inputs and outputs of the program. We are free to do this in any way we choose (the isomorphism theorem we will look at shortly applies to *any* input-output interpretation of the program). It is possible that some of the possible input-output assignments will result in the same f_x, so f_x could be considered a set corresponding to all the input-output assignments that correspond to the same function. We can also say that f_x is a member of the set X_a of all partial computable functions that can be represented by X_a under an appropriate interpretation (input-output assignment).

If we are to assume that these programs are objective entities (as in, countables for our probability theory), then they must not be mere artifacts of our choice of language *or* our choice of translation manual (a "translation manual" being any set of rules that tells us how to translate programs in one language into programs in another). For instance, suppose we find that when we translate the same program X_a back into language A, that we have the following situation:

$$X_a \longrightarrow_\gamma I_b \longrightarrow_\theta Y_a \tag{4.64}$$

then we have a problem, since the identity of the program seems to depend on our translation scheme. Even if we required that a "translation manual" be "two-way" (required to act in both

directions between two languages in a consistent fashion) we might still end up with

$$X_a \longrightarrow_\theta I_b \longrightarrow_\gamma Q_c \longrightarrow_\nu Y_a \tag{4.65}$$

which gives us the same problem of identity for programs. And even if we could show that the above could never happen, we still might end up with

$$\begin{aligned} X_a &\longrightarrow_\theta I_b \\ X_a &\longrightarrow_\gamma J_b \end{aligned} \tag{4.66}$$

which yields the same identity problem, but due now to an inherent ambiguity of translation.

It is this kind of identity problem that lead Quine to adopt his ontological relativism [166]. In the extreme case, one might wonder whether a suitable choice of translation manual might allow *any* program to be mapped onto any other program, especially if we allow any kind of loopy and strange translation scheme. And Kleene's theorems seem *prima facie* to strengthen this worry, since they imply both program redundancy and ambiguity.

Theorem 4.37. ***Program redundancy***: *there are multiple programs that represent the same partial computable function:*

$$\forall X_a, f_x \in \{X_a\}, \exists Y_a : (X_a \neq Y_a) \wedge (f_x \in \{Y_a\}) \tag{4.67}$$

Proof. Since there exists a UTM—call it $UTM(X_a)$—that can simulate program X_a, there are necessarily multiple programs that can represent any function that X_a can represent (trivially, at least X_a and $UTM(X_a)$). □

Theorem 4.38. ***Program ambiguity***: *a single program can represent multiple partial computable functions:*

$$\forall X_a, \exists f_x, g_x : (f_x \neq g_x) \wedge (f_x, g_x \in \{X_a\}) \tag{4.68}$$

Proof. This follows simply from noting that when $UTM(X_a)$ simulates program X_a, it can be interpreted both as the functions $UTM(X_a)$ and as $f_x(k)$. □

The existence of both program redundancy *and* ambiguity seems to suggest, on the face of it, that we *do* have a program identity problem. However, this might still turn out to be perfectly benign. So long as the set of partial functions that can be assigned to a program remains the same after translation, our programs can still be viewed as objective objects of human thought. Granted, one program can represent multiple functions, but singling out one function as *the* function that the program computes is always just a synthetic interpretation, anyway, so this is not a problem

for our ontology. And granted, one function can be represented by multiple programs, but this is not a real problem for program identity, so long as the same program, after translation, represents the same functions. Given that functional interpretations of programs have had their usefulness very well established, if we found that these interpretations did not survive translation, then this fact, combined with program redundancy and ambiguity, would create an insurmountable program identity problem. For instance, we might have

$$\begin{aligned} X_a &\longrightarrow_\theta I_b \\ X_a &\longrightarrow_\gamma J_b \end{aligned} \tag{4.69}$$

where

$$\begin{aligned} X_a &= \{f_x, g_x\} \\ I_b &= \{f_x, h_x\} \\ J_b &= \{h_x, g_x\} \end{aligned} \tag{4.70}$$

Any interpretation that does not survive translation is purely synthetic, and we can therefore eliminate its function(s) from the set of functions represented by the program. We will write "$X \longrightarrow Y : \{f_1, f_2, \ldots\}$" to mean that $f_1, f_2, \ldots$ are the only partial functions that are *objectively* and determinately represented by X given that it can be translated to Y. (Note that the recursion theorem allows us to "translate" a program into another program in the same language.)

$$\begin{aligned} X_a &\longrightarrow_\theta I_b : \{f_x\} \\ I_b &\longrightarrow_\mu J_b : \{h_x\} \\ J_b &\longrightarrow_\gamma X_a : \{g_x\} \\ X_a \longrightarrow_\theta I_b \longrightarrow_\mu J_b &\longrightarrow_\gamma X_a : \{\} \end{aligned} \tag{4.71}$$

Here, there are no longer any partial functions represented by X_a. We have completely "deconstructed" X_a, so that its meaning has literally disappeared into thin air.

Rogers' isomorphism theorem shows us, however, that this particular kind of deconstruction, at least, is not possible (at least not if we interpret recursive structures as partial computable functions).

Theorem 4.39. ***Rogers' isomorphism theorem****: there is an isomorphism (one-to-one mapping) between any two enumerations of the partial computable functions. In other words, given any mapping between enumerations (Gödel numberings) of programs in any number of different Turing-complete languages, the partial computable functions that are represented by the programs are invariant.* [178, p.338]

Thus, even though there is inherent redundancy and ambiguity as to which partial computable functions are represented by the different programs, a given program is characterized by the *same* set of partial functions across any number of Turing-complete translations. Quinean arguments for relativism do not seem to hold up under the weight of this result, so long as our language is Turing-complete, and so long as we accept the Church-Turing thesis.

Hence, so long as our translation manual does a proper translation of any partial function in one Turing-complete language into a partial function in another Turing-complete language, this theorem guarantees that we will end up with a functionally equivalent set of partial functions, no matter how strange and loopy and intentionally devious our method of translation. We can make a good case, on the basis of this result—and pursuant to our acceptance of the Church-Turing Thesis—that there exist truly invariant concepts of the human mind, corresponding to the logico-mathematical objects of recursion theory (the recursive structures). And even if we do not accept Church-Turing, Rogers' theorem is still a very powerful result within the context of recursion theory, giving Turing-complete languages a degree of objectivity that they would not otherwise have. The "non-deconstuctibility" demonstrated by the theorem is not absolute, however, since it applies only to *functional* interpretations of recursive structures, and I have already argued that programs are only optionally interpreted as functions. However, it is arguable that any substantive application or interpretation of a recursive structure ought to be conceivable in functional terms, so it is difficult to see this as a real limitation. In any case, it is unlikely that one could ever *prove* that one had a scheme that exhausted all possible interpretations, and functionalism seems as universal an overarching interpretative framework as we have at the moment. Sometimes I may prefer to talk about algorithms, rather than functions, but algorithms are straightforwardly subsumable under functions.

4.3.7.3 Algorithmic Probability

Algorithmic probability can be developed as an extension to, or refinement of, generative probability. Recall that generative probability asks us to count, in the classical manner, the number of observer-generating objective (ontic) entities. In an algorithmic context, this will mean the number of observer-generating algorithms, or just "observer-algorithms", for short.

So how do we count observer-algorithms? First, we need to note that an algorithm that generates an observer is still just another program, not a fundamentally different kind of entity. Thus, when we say that we are "counting" observer-algorithms, we are counting *programs* that fit into the *category* of an observer-algorithm. In other words, we are still counting *programs* in the denominator, while counting algorithms (programs grouped into categories) in the numerator. It is all still program

counting, and the ontic entities are still programs, *not* algorithms. This is an important point, as it is not possible to defend the use of observer-algorithms *per se* as ontic entities. It is the set of programs that is maximally analytically expressive. The concept of "algorithm", on the other hand, refers to the production of a result in a finite number of program steps (or else a convergence on some result in the limit). This requires the idea of program "output", which requires an arbitrary, synthetic split of the program into "output" and "internals". This split has no analytic justification, is not a feature that is retained across Turing-complete languages. The same program can be interpreted as outputting different results, and so an "algorithm" is a way of categorizing our ontic entities (programs), not a kind of ontic entity itself.

According to algorithmic information theory, to count programs, we use the following simple equation for probability, that we derived earlier from Kraft's inequality:

$$p(x) = b^{-H(x)} \tag{4.72}$$

where b is the numerical base of the program code, x is the program, and $H(x)$ is the algorithmic information content of x. Without loss of generality, we will assume a binary base ($b = 2$) for convenience:

$$p(x) = 2^{-H(x)} \tag{4.73}$$

(In the context of ASU, x will typically be a conscious mental state, or more precisely a program that acts as an algorithm for producing that mental state.)

Recall that $H(x)$ can be understood as the least number of bits it takes to write an algorithm to output x. This measure, $P()$, may not seem at first to be the result of a literal "count" of programs, but it can be understood that way.

We start by calling the set of all programs $\mathcal{P}$. We enumerate these programs,

$$\mathcal{P} = \{P_k : k = bL + c = 1, 2 \cdots\} = \{P_L^c\} \tag{4.74}$$

where

$L = L(P_k) = |P_k| = \frac{k-c}{b}$ is the bit-length of program P_k, and
$c =$ index to $\mathcal{P}_L = \left\{P_L^0 \cdots P_L^c \cdots P_L^b\right\} \subseteq \mathcal{P}$, the subset of programs in $\mathcal{P}$ of length L.

The subset of $\mathcal{P}$ limited to bit-lengths $L \leq n$ is:

$$\mathcal{P}^n \tag{4.75}$$

Consider programs of size n, so there are b^n programs we could write in this space. A simple principle of program indifference would tell us that the probability of the k-th program in a random pick of

programs should be, by the classical probability ratio,

$$p_k = p(P_k) = \frac{1}{b^n}? \tag{4.76}$$

But this is as far as we can go if we are talking about straight program counting (in the denominator). The problem here is that we run up against the problem of infinite sample spaces, as we cannot generalize this result to programs of arbitrary size, since then we get

$$p_k = \lim_{n\to\infty} \frac{1}{b^n} = 0? \tag{4.77}$$

and we cannot talk about the probability of picking a program if the probability of all programs is zero. By doing straight program-counting, all programs are equally probable, and since the sample space is infinite, $p(k)$ would be undefined for the same reasons we said that $p(3)$ was undefined.

But imagine that it only takes L bits for P_k to do its work (generate the required output), no matter its actual length. This means we can fill up the remaining $n - L$ bits with whatever random or useless nonsense we want, and the program will function identically (assuming, of course, that our programs are written in a prefix-free code). This means that there are really 2^{n-L} different versions of this program in our set of 2^n programs. They are distinguishable as programs, but are necessarily identical as algorithms. Hence, we will generally talk about this set of programs (or algorithm) *as if* it were a single program. We can now find the probability of picking one of these programs at random from the set of all programs of length n, by using the classical probability ratio:

$$p_k = \frac{b^{n-L}}{b^n} = b^{n-L-n} = b^{-L} = \frac{1}{b^L} \tag{4.78}$$

where L =the length of the program.

Notice that n has now cancelled out of the equation, thus side-stepping the problem of infinity. Note also that this is a *different* probability than the $p(k)$ candidate we considered earlier, since technically we counted more than one program code. Hence, we could consider this to be a very simple and trivial form of algorithm-counting. However, given that straight program-code counting is undefinable, and given that the program codes we are grouping together here are very trivially identical, given the assumption of a prefix-free code (and we can always choose to make our code prefix-free), it makes sense to consider the above to be "standard" program counting, and generally reserve "algorithm" for more interesting, nontrivial groupings of programs.

Next, we wish to consider the subset $\mathcal{P}(m)$ of $\mathcal{P}$ of programs that generate m (later, m will be the conscious mental state of an observer, but for now, it could be any kind of data at all). When we say m is the "output" of program P_k, what we will mean more precisely is that m is itself a program that tests other programs for whether they generate the output in question. This allows

more general objects to be considered as successful outputs, other than just those of *particular* bit sequences (and this flexibility will be required to deal with the case of conscious mental states, for instance). So we define

$$m(k) = \begin{cases} 1 & \text{if } P_k \text{ generates } m \\ 0 & \text{otherwise} \end{cases} \tag{4.79}$$

and membership in $\mathcal{P}(m)$ is defined as

$$\mathcal{P}(m) = \{P_k : m(k) = 1\} \tag{4.80}$$

In general, of course, there could be many programs that generate m—in fact, given Kleene's recursion theorems, there will be a countable infinity of them. We enumerate these programs, yielding the set of all programs, $\mathcal{P}(m)$, that produce output m:

$$\mathcal{P}(m) = \{P_i(m)\} \tag{4.81}$$

so that the $P_i(m)$ are subject to the same partial ordering as the P_k. In this case, however, there is no longer any simple relationship between i and b, L, c (since there is no simple way, other than by actual enumeration, to know how the programs that generate m are distributed in $\mathcal{P}$).

Again, to restrict ourselves to programs of a bounded size n we use

$$\mathcal{P}^n(m) = \{P_k : |P_k| \leq n \wedge m(k) = 1\} \tag{4.82}$$

We can now express the probability $p^n(m)$, in a random selection from the set of all programs up to size n, of picking one that outputs m, by using the same classical ratio we used for single programs, but generalized to $\mathcal{P}^n(m)$:

$$\begin{aligned} p^n(m) &= \sum_{P_k \in \mathcal{P}^n(m)} p_k \\ &= \sum_i b^{-|P_i^n(m)|} \end{aligned} \tag{4.83}$$

which is only strictly defined for finite n, but which we can generalize to $\mathcal{P}(m)$ as

$$\begin{aligned} p(m) &= \lim_{n\to\infty} p^n(m) \\ &= \lim_{n\to\infty} \sum_i b^{-|P_i^n(m)|} \end{aligned} \tag{4.84}$$

so long as it converges on a well-behaved (finite and generally non-zero) result.

Definition 4.40. The enumeration of the elements of $\mathcal{P}^n(m)$ we will call the "Solomonoff sequence". The summation of the members of the Solomonoff sequence from (4.84) will be called the "Solomonoff series", and $p_S(m)$ will be called the "Solomonoff probability" (we can drop the S when no confusion results).

And in fact, Solomonoff [205] proves that the Solomonoff series does converge, and does so very quickly.

Theorem 4.41. ***Solomonoff's convergence theorem****: the Solomonoff series (yielding $p(m)$ in the limit) converges (and does so exponentially).*

Proof. (Sketch) We can see in outline why this theorem follows by looking at the exponential decrease in the series of probabilities in (4.84). Since the programs in $\mathcal{P}^n(m)$ are partial-ordered by length, and since probability falls off exponentially to zero with program length, $p^n(m)$ will decrease exponentially to zero (see [205] for a full proof). □

Given that the infinite limit does exist, Solomonoff writes the formula for $p(m)$ more compactly as

$$p_S(m) = \sum_{i=1}^{\infty} b^{-|P_i(m)|} \tag{4.85}$$

with marginal probabilities

$$p(P_i(m)) = b^{-|P_i(m)|} \tag{4.86}$$

If we gloss over the distinction between i and $P_i(m)$, and between m and $\mathcal{P}(m)$, where no confusion will result, we can write the very compact and intuitive version

$$\begin{aligned} p_S(m) &= \sum_{i \in m} p(i) \\ &= \sum_{i \in m} b^{-|i|} \end{aligned} \tag{4.87}$$

where

$$p(i) = b^{-|i|} \tag{4.88}$$

Definition 4.42. Solomonoff complexity, information, or entropy, $H_S(m)$, is simply the inverse of Solomonoff probability:

$$H_S(m) = -\log_b p_S(m) \tag{4.89}$$

Given that the terms in the series do decrease exponentially, it follows that the shorter (*i.e.,* earlier) programs will contribute by far the most to the probability. Hence, we can cut off the summation at a relatively low i, and still have a good chance of a reasonable approximation to the actual probability. In the most extreme case, we could simply take the *first* term in the sum and forget the rest, since at least we are guaranteed that no other programs will be *more* probable than this one—although there still may be some, even many, that are *equally* probable, since there is only a *partial* ordering on the programs (thus, as one enumerates the terms in the Solomonoff series, it follows that before p_S decreases, one may have to go through as many as $b^L - 1$ other

equally probable cases). Therefore, taking only the first term in the summation is still potentially only a rough approximation to the actual Solomonoff probability, and may produce significantly different results in some situations (although in other situation, it may serve virtually as well). For most applications, it seems in practice that the two measures are interchangeable, and algorithmic information theorists tend to virtually equate them. I will do the same in some of my arguments, as it is much simpler to think in terms of the first-term approximation. The possibility that we may be overlooking a reason to consider the full Solomonoff series should not be forgotten in this, however.

Definition 4.43. The "Kolmogorov probability", $p_K(m)$, is the first-term approximation to the Solomonoff probability:

$$p_K(m) = 2^{-H_K(m)} \tag{4.90}$$

where $H_K(m)$ is the "Kolmogorov complexity", or the length of the shortest program that generates output m:

$$H_K(m) = \min_i (|P_i(m)|) \tag{4.91}$$

Keep in mind that "the shortest program" here does *not* mean "the unique shortest", since there could be as many as $2^{H_K(m)}$ binary-coded programs in total that all generate m.

Of course, we could also create any number of other approximations, by cutting off i at other threshold points. Perhaps most sensible is to cut it off at a certain program size, rather than a certain value of i.

Definition 4.44. Cutting the Solomonoff series off at program length L (so at maximum $i = L$) will yield an "L-order approximation" of the Solomonoff probability.

Note that a first-order approximation takes into account all programs of minimal length that generate m. The Kolmogorov approximation is worse than this, as it only considers the *first* program. In a first-order approximation, we are at least assured that all programs we ignore are at least an order of magnitude less probable than the ones we are using. With the Kolmogorov approximation, we could be ignoring a great many programs that are actually just as probable as our chosen program. First-term and first-order approximations are most useful when we just want to get a complexity or entropy measure for the original data m. In this case, it does not matter so much exactly which programs contributed to the measure. However, in the case where we want to actually calculate the algorithmic probability of a *particular* encoding of m, as opposed to any others, we may need the Solomonoff measure. The Kolmogorov measure only cares about two encodings: the original and the compressed versions. Solomonoff's measure takes into account *all* encodings that contribute to the overall information content of m. Hence, we can use it to calculate (conditional) probabilities of particular encodings (particular members of $\mathcal{P}(m)$), given the original encoding for m.

Definition 4.45. The probability of the i-th member of $\mathcal{P}(m)$, $P_i(m)$, is

$$\begin{aligned} p_S(P_i(m)) &= \frac{b^{|P_i(m)|}}{b^{H_S(m)}} \\ &= b^{|P_i(m)|-H_S(m)} \end{aligned} \tag{4.92}$$

where $|P_i(m)|$ is the marginal entropy of a particular encoding of m, independent of its being an encoding, efficient or otherwise, of m. Hence we will abbreviate it as $L(i)$, and write the above more compactly as

$$p_S(i \in m) = b^{L(i)-H_S(m)} \tag{4.93}$$

Note that Solomonoff complexity $H_S(m)$ computes the *average* length of an m-generating program, as opposed to the Kolmogorov complexity, $H_K(m)$ which computes the *shortest* length.

If we wish to specify that we are using a particular Turing-complete language, $\mathcal{L}$, we can use superscripts, as in $H^{\mathcal{L}}(m)$ and $p^{\mathcal{L}}(m)$. The particular language will often be omitted for ease of reading (I have already given my reasons for preferring the SK-calculus, so the reader can generally assume, in this dissertation, that $H(m)$ means $H_S^{SK}(m)$, the Solomonoff complexity in the SK-calculus).

Theorem 4.46. ***Solomonoff uncomputability****: the Solomonoff probability $p_S(m)$, as well as any n-order or n-term approximation to p_S, for any natural number n, is uncomputable.* [204]

Proof. (Sketch) The theorem follows if we require that any valid compression of a program must eventually halt, and from the fact that the halting problem (the problem of deciding whether a given program will halt) is undecidable [217]. However, even if we do not expect an m-generating program to halt, the problem of deciding whether a program will produce a given program is likewise undecidable, for the same reasons the halting problem is (in fact, the halting problem is actually just a special case of the output-generation problem). Therefore, if we enumerate all programs, and run each one, as we go along—in order to see if it produces m—then for those programs that *do* generate m, we will eventually *know* that they do. However, for those that do *not*, we will never know that they do not, since to know this would, in general, imply that we had solved the halting (or output generation) problem. Hence, $\mathcal{P}(m)$ and $p_S(m)$ are uncomputable, even to a first (or higher) term/order approximation. □

While it may be disappointing that we cannot reliably compute algorithmic probabilities, it should not cause us undue consternation from an ontological point of view. The main point is merely that there *are* such probabilities, and that they are objective—not that we necessarily need to be able to compute their values for ourselves.

However, while n-order and n-term approximations are uncomputable, another kind of approximation *is* computable:

Definition 4.47. The "t-step" n-order (or n-term) approximation computes an approximation to the Solomonoff series by enumerating each program, P_k, checking to see if it generates m within t program steps, and if it does, deciding that $P_k \in \mathcal{P}(m)$.

Note that if a t-step approximation is not *also* either n-order or n-term for finite n, then the measure will still be uncomputable.

Theorem 4.48. ***Solomonoff limit-computability:*** *the Solomonoff probability $p_S(m)$ is limit-computable if m is limit-computable; and more specifically, it is $(n+2)$-limit-computable if m is n-limit-computable.*

Proof. Assume m is computable. The Solomonoff series can then be limit-computed in the process of 1-limit-computably enumerating $\mathcal{P}(m)$. Therefore, the Solomonoff series, and hence Solomonoff probability, is 2-limit-computable. Dropping the assumption that m is computable, if m is n-limit-computable, the Solomonoff probability is $(n+2)$-limit-computable. Hence, so long as m is limit-computable, so is the Solomonoff probability. □

A t-step n-order approximation of Solomonoff probability approaches the full limit-computation as t and n approach infinity.

The convergence theorem of Solomonoff's gives us reason to believe our attempts to compute approximations should be reasonable. Without the convergence theorem, we might worry about the practicality of even approximating Solomonoff probabilities. While this would not likely have serious ontological consequences, its epistemological consequences could feasibly be significant, since falsifiability might depend on our ability to at least compute reasonable estimates (for instance, I will argue later that the falsifiability of the algorithmic interpretation of quantum probability may well turn on our ability to make reasonable estimates of the information content of conscious states and/or of the early universe). The convergence theorem makes it more likely that such estimates will be practical.

The more serious issue for Solomonoff, from an ontological perspective, is that both of the above two results, while apparently speaking very favourably to the objectivity of algorithmic probability, can still only be stated *relative* to a particular language. But, surely, our choice of language should not affect $p_S(m)$ at all, if it is to be taken as an objective measure. Solomonoff himself has said [206] that he spent years searching for a theorem that could show that his probabilities were completely objective and *a priori*, by proving that the choice of language does not ultimately matter. He was

largely (or at least partly) successful, but not entirely so. However, he has since said that just when he thought he had accomplished the feat, he decided it did not matter, and that he had no need for objectivity. His reason for not requiring objectivity, in the end, was essentially that one's choice of language can simply be considered a model of one's prior knowledge (or beliefs), taking probabilities to be epistemic (or doxastic). Indeed, one can assume that—if the human brain can be modelled analytically—that there is a combinator that represents any given human brain, and that this combinator can be used as a model of prior knowledge and/or belief, in the construction of a combinator that computes $p_S(m)$, either in the limit or as an approximation.

So, while the purely analytic nature of algorithmic probability may seem superficially counter to the subjectivist nature of Bayesianism, Solomonoff probabilities can actually be quite a comfortable fit with at least certain flavours of Bayesianism (and it would appear that Solomonoff himself was some variety or other of Bayesian, at least in later years).

Our main issue here, however, is that we do not want to be stuck with either epistemic *or* doxastic probabilities, in the first place. We are looking for a model of *objective chances*. Thus, we want our choice of language to be as free from *any* model of the observer's beliefs or knowledge as possible. This is why we spent so much time choosing which analytic languages were formally and conceptually the simplest, since to justify our choice of language, we will at least have recourse to Occam's razor, even if we cannot actually *prove* that certain languages are superior to others.

Question. *Given that we accept, for reasons of simplicity, that there are only certain languages that are reasonable candidates for an analytic basis language, can Solomonoff probability serve as a sufficiently objective measure for the purposes of Everettian quantum mechanics?*

Definition 4.49. Define $p_s^a(x)$ as the Solomonoff probability, calculated using language a.

Ideally, we would like to be able to show that

$$\forall a, b, x : p_S^a(x) = p_S^b(x) \tag{4.94}$$

so that our probabilities are independent of our choice of language. However, this is clearly not going to be the case for arbitrary a and b, since a highly contrived language, or one replete with priori knowledge, will very clearly produce much higher probabilities for those structures that share high mutual information with this built-in knowledge base. However, we have already decided that we will invoke Occam's razor to limit a and b to reasonably "simple" languages. However, as we have seen, this still does not result in a *single* language. It is still debatable, for instance, which is the simplest of the λ-calculus or SK-calculus. True, I have already argued why I tend to favour the SK-calculus, but even so, this is a largely aesthetic judgement, and there are almost certainly other

candidates that would be competitive with combinatory logic, that I have not even considered (and that may not have even been thought of yet). Thus, if we define $\mathcal{A}$ informally as the set of all such acceptable analytic basis languages, a more practical and perhaps achievable goal would be to show that

$$\forall x, \forall a, b \in \mathcal{A} : p_S^a(x) = \varepsilon(a, b, x) p_S^b(x) \tag{4.95}$$

where $\varepsilon(a, b, x)$ is close to unity. Or—virtually the same thing—we could show that

$$\forall x, \forall a, b \in \mathcal{A} : H^a(x) = H^b(x) + \epsilon(a, b, x) \tag{4.96}$$

where $\epsilon(a, b, x)$ is sufficiently small. I have no formal requirements for $\epsilon()$, and how small it needs to be. Ideally, we would like it to be zero, and perhaps there are some who will accept nothing short of that. However, if we could show that it was very small, it would be difficult to argue that Solomonoff probabilities were entirely subjective. There is no absolute line to draw here, because "objectivity" does not have to be an absolute property; it can admit of degrees. On the other hand, for Everettian probabilities to be objective in the sense we want them to be, there may not be a lot of leeway here. Given the accuracy with which quantum probabilities are calculated, and the accuracy of the predictions made with them, we basically need to show that $|\epsilon(a, b, x)|$ is either zero, or vanishingly small, if we are to argue that quantum probabilities are both algorithmic and in any way fundamentally objective in nature.

As it turns out, $\epsilon(a, b, x)$ is quite strongly bounded, and very close to zero for large programs.

Definition 4.50. Given arbitrary Turing-complete languages a and b, we construct a "translation manual from a to b", defined as a program $T_b^a(P_i^a)$ in language a that takes the i-th program P_i^a in language a, and produces an equivalent program P_j^b, the j-th program in language b, such that there is a one-to-one correspondence between programs in a and those in b. Our ability to construct the manual follows from Kleene's recursion theorems and the existence of UTMs. (Note that T_b^a denotes an arbitrary such translation manual; it is not unique.)

Theorem 4.51. ***Solomonoff invariance:*** *Solomonoff complexity, $H_S(m)$, is invariant across Turing-complete languages, up to an additive constant. Thus, as m gets larger, the dependence of $p_S(m)$ on the choice of language approaches zero.*

$$\begin{aligned} \forall m, \forall a, b : H_S^a(m) &= H_S^b(m) + \epsilon(a, b) \\ H_S^a(m) &= H_S^b(m) + O(1) \end{aligned} \tag{4.97}$$

So that

$$\lim_{m \to \infty} \left(H_S^a(m) - H_S^b(x) \right) = 0 \tag{4.98}$$

And for sufficiently large m,

$$p_S^a(m) \approx p_S^b(m) \tag{4.99}$$

Proof. (Sketch) See [204] for a complete proof. I will prove it here for the simpler case of the first-term Kolmogorov approximation to Solomonoff probability. If P_i^a is an arbitrary program in language a, with $H_K(P_i^a)$ and $P_{i'}^a$ the corresponding optimal-length program in a, then if $P_j^b = T_b^a\,(P_i^a)$ is a corresponding program in language b, the length of P_j^b is

$$L(P_j^b) = L(P_i^a) + L(T_b^a) \tag{4.100}$$

and the optimal length for the language b program is therefore

$$H_K(P_j^b) \leq H_K(P_i^a) + L(T_b^a) \tag{4.101}$$

so H_K in b is equal to that in a, plus an amount no more than the length of the translation manual $L(T_b^a)$, which is a constant for the given languages, independent of P_i^a and P_j^b. □

Given Solomonoff invariance, if we restrict ourselves to members of $\mathcal{A}$, the magnitude of $\epsilon(a,b)$ will not only stay constant as m grows, but will actually be quite small in absolute terms, as well[45], so long as we restrict ourselves to $\mathcal{A}$. Short of finding that $|\epsilon(a,b,x)| = 0$, this is the strongest result for objectivity that we could expect to find. It shows, not that our measure is absolutely language-independent, but that it approaches language independence for large x and is nearly language-independent for any x that is sufficiently large in comparison to the maximum translation overhead between any two languages in $\mathcal{A}$. This seems indicative of an underlying objective content to the Solomonoff measure, with the remaining $O(1)$ overhead representing a limitation in our ability to represent this content distinctly, rather than a limitation on the objectivity of that content itself. Admittedly, however, I cannot prove this. There may still be those who hold out for the allowance of very contrived languages, which, although $O(1)$, are so large that this huge "overhead" will dominate the Solomonoff measure, making it clearly subjective.

Such subjectivity is fine in general—Solomonoff himself was a Bayesian subjectivist. For epistemic or doxastic probabilities, extremely complex languages are permitted, since they serve as models of the belief system of the observer, transforming algorithmic probabilities into subjective

[45] And certainly the translation overhead involved in describing an m that represents a human conscious state should be relatively tiny. There is, however, a possible catch here: by claiming that the translation overhead will probably be tiny compared to the information content of a human consciousness, I am assuming that the latter is itself *not* tiny, in spite of the fact that a primary feature of the algorithmic interpretation of quantum probabilities will be the highly compressible nature of human consciousness. At this point in time, it is not really possible to say exactly *how* compressible a human consciousness really is, but it still seems far-fetched to think that it would contain a number of bits as few as those in a translation manual between the λ and SK-calculi (or similarly minimal analytic languages). This is, for the time being, however, admittedly nothing more than intuitive presupposition on my part.

Bayesian priors. However, we are interested here in *objective* chances—ontic probabilities—so we must necessarily restrict ourselves to $\mathcal{A}$, and it seems that the invariance and isomorphism theorems give us reason to believe that this practice is well-grounded.

Together, the invariance and isomorphism theorems provide a strong rational grounding for algorithmic probabilities. The invariance theorem demonstrates language-independence (up to an additive constant), while the isomorphism theorem (assuming Analytic Church-Turing) demonstrates objectivity, at least for functional and algorithmic interpretations.

4.3.8 Conclusions

I will not claim that Solomonoff (algorithmic) probability is better than all other interpretations of probability, for all cases (it may or may not be). There may be cases where trying to find an algorithmically-justified prior probability is a fool's game (even if one believes that such a prior exists, it may be thoroughly impractical to even approximate it). And while Solomonoff probabilities may be useful in working with epistemic and doxastic probabilities, and in Bayesian inferencing (as they promise a relatively well-justified notion of prior probability), there may be objections that can be raised to such a project, and these issues are beyond the scope of our concerns here.

What I hope I *have* established is that algorithmic probability provides our best foundation for an objective *a priori* probability theory, which is necessary if we are to make an informed attempt to calculate Everettian quantum probabilities *as objective chances*. I have argued, counter to much common wisdom, why metaphysically-grounded generative probabilities are the best approach to a theory of (open) single-case chance, and further why algorithmic probabilities are the best-justified, most general way to implement objective generative single-case probabilities. Accepting this conclusion has required that we accept a number of assumptions, such as the analytic completeness of Turing machines, the *a priori* simplicity of certain Turing-complete languages (members of $\mathcal{A}$), and that $\epsilon(a, b)$ is (probably) sufficiently small to allow us to consider algorithmic probabilities to be invariant (across members of $\mathcal{A}$). These assumptions, while not proven, have compelling rational motivations, and are compatible with the assumptions Everett is already asking of us with his relative state interpretation—and I believe they are at least rationally compelling enough, that if one could prove the Born rule from *a priori* grounds, given these assumptions, that even the hardest-core Born rule objector would have to admit that our *a priori* grounds are far superior to the previous standard that they themselves have generally upheld: that of branch-counting.

In any case, I hope the argument for an algorithmic approach to probability is compelling enough that even the most sceptical reader will not dismiss it out of hand, and will be interested in exploring with me where such an approach may lead.

5 Self-location and Self-selection

5.1 Self-location

Anthropic, or self-selection, principles will form a foundation for much of what is to come in later chapters. However, there are many traps that are laid for the unwary who is not armed with the ability to choose the appropriate indifference principle. If we proceed on intuition alone, without an understanding of the deep metaphysical and epistemological differences that separate different views on these matters, then we are doomed.

Before we proceed further, we need to take a thorough look at the different philosophical views of probability that have lead to different views on the anthropic principle, and likewise many-worlds. The Sleeping Beauty puzzle, a recent paradox that has garnered much attention in the literature in the last ten years, is an ideal platform for exploring these foundational issues, as it encapsulates most of what is essential about "self-locating belief": how beliefs and probabilities function when the subject does not have certain knowledge about where (or when) they are in the universe. The issues that arise from this will lead us from self-location right into (anthropic) self-selection. We will discover that knowing how to deal with this one probability puzzle in terms of generative probability theory will lead us through a lot of the territory necessary to deal with the issue of the Born rule objection.

The Sleeping Beauty puzzle was introduced into the literature by Adam Elga [75], whose solution was criticized by David Lewis [130], resulting in a protracted debate between supporters of the two proposed solutions. No clear consensus has emerged (although it does seem that Elga's solution currently has an edge in popularity).

5.1.1 The Sleeping Beauty Puzzle

5.1.1.1 The Puzzle

Sleeping Beauty has agreed to participate in an experiment in a sleep lab. She is given the following details:

On Sunday night, she will settle down to sleep. Once she is asleep, the experimenters will toss a coin. Two outcomes are possible:

1. *The coin is heads.* In this scenario, Sleeping Beauty wakes up on Monday morning; she is interviewed, and then she goes home.
2. *The coin is tails.* In this scenario, Sleeping Beauty wakes up on Monday morning; she is interviewed, after which she is given a drug that erases her memory of having woken up on Monday, or anything else that happened on Monday. She goes back to sleep and awakens on Tuesday morning, whereupon she is interviewed again. At the time of the interview, she has no knowledge of whether it is Monday or Tuesday. Then she goes home.

When Sleeping Beauty is interviewed, whether it is Monday or Tuesday, she is asked, "What is the probability that the coin landed heads?"

5.1.1.2 The Two Solutions

How Sleeping Beauty should rationally answer this question is still a matter of debate, as no clear consensus has emerged thus far. Two possible solutions both seem quite plausible:

1. The "Thirder" solution: the probability of heads is 1/3.
   There are three equally possible states for Sleeping Beauty to be in at the time of the interview. First, it could be Monday, and the coin flip could have resulted in heads. Second, it could be Monday, and the coin flip could have resulted in tails. Third, it could be Tuesday, after Monday's coin flip resulted in tails. *All three states are subjectively indistinguishable to Beauty*, so she must assign them equal credence. Since only one of the three states has a result of heads, the probability of heads must be 1/3 .
2. The "Halfer" solution: the probability of heads is 1/2.
   It is clear that there is a 50% chance that the coin landed heads on Sunday night, so how could the probability be any different on Monday morning, given that *Beauty has received no new information at the time of the interview*? After all, everything on Monday morning is exactly as she expected it to be: she knows no more about the coin flip than she did the night before. Hence, the probability of heads must be the same as it was on Sunday, or 1/2.

5.1.1.3 Centred and Uncentred Worlds

Definition 5.1. An *uncentred world* is a (possible or actual) world, without any specified observer (although it may still incidentally have observers in it).

Definition 5.2. A *centred world* is a (possible or actual) world, along with a specified observer within that world (who may be specified in any number of ways; typically, a space-time location is specified). We say that the world is *centred on* that observer.[46] Note here that by *observer* is meant

[46] The distinction between centred and uncentred worlds plays a central, and similar, role in the absolute idealism of Bradley [27, 170] as well as the modal realism of David Lewis [129] (both are *a priori* many-worlds cosmologies). Bradley calls the observer the "finite centre" that selects out one world from the infinite manifold of possible worlds—basically a nineteenth century precursor to the anthropic principle and self-selection effects in general.

a particular conscious mental state, not an entire history of such states, which would constitute a particular *person*.

While the idea of a "centre" is similar to algorithmic synthetic unity, it entails somewhat different philosophical commitments, and so it would be very confusing to confound the two, and I will keep the two terminologies separate. Nonetheless, one can roughly compare the centre of a world to the conscious state around which a world is defined in terms of algorithmic synthetic unity (again, this is a rough analog, not an equivalence).

In the Sleeping Beauty puzzle, the assignment of probabilities will turn out to crucially depend on the distinction between *centred* and *uncentred* information.

Definition 5.3. *Centred information* is information received about a centred world by the observer that the world is centred on.

Definition 5.4. *Uncentred information* is information about an uncentred world.

Centred information, then, is inherently *synthetic*.

5.1.1.4 Preliminaries

I will mostly adopt Elga's notation, since it is used throughout the literature. We will use these abbreviations:

0: Sunday
1: Monday
2: Tuesday

H: Heads
T: Tails

which can be combined to describe five possible situations Beauty can find herself in:

0: Sunday (no coin flip)
$H1$: Monday, heads
$H2$: Tuesday, heads
$T1$: Monday, tails
$T2$: Tuesday, tails

Lewis summarizes the common ground between Thirders and Halfers as follows:

1. When Beauty awakens there are three centred possible worlds: $H1$, $T1$, $T2$.
2. The probabilities for tails on Monday and Tuesday are the same:

$$p(T1) = p(T2) \tag{5.1}$$

3. When Beauty is told it is Monday, her credence function changes from $p()$ to $p_+()$:

$$p_+(H) = p(H|\neg T2) = p(H|H1 \vee T1)p_+(T) = \ldots \tag{5.2}$$

4. Since for $p()$, H iff $H1$:

$$\begin{gathered} p(H) = p(H1)p(T) = \\ p(T1 \vee T2) = p(T1) + p(T2)Pp(H|H1 \vee T1) = \\ p(H1|H1 \vee T1) = \frac{p(H1)}{p(H1) + p(T1)}p(T|H1 \vee T1) = \\ p(T1 \vee T2|H1 \vee T1) = \frac{p(T1)}{p(H1) + p(T1)} \end{gathered} \tag{5.3}$$

5. On Sunday night, if we call Beauty's credence function $p_-()$:

$$p_-(H) = p_-(T) = \frac{1}{2} \tag{5.4}$$

6. On Monday there is no new uncentred information.

7. The centred information Beauty gets is *only*

$$\begin{gathered} H1 \vee T1 \vee T2 = \\ \neg H2 \end{gathered} \tag{5.5}$$

5.1.1.5 Elga and the Thirders

Elga's position is based on his principle of indifference for self-locating belief, expressed in [76], which I will paraphrase as:

Principle 5.5. *Elga's Self-locating Principle of Indifference: Subjectively indistinguishable centred worlds, that share the same uncentred world, should be assigned the same credence.*

Elga distinguishes this from a slight variation that he calls his "absurd-principle-of-indifference-that-I-do-not-support" [76], which I will paraphrase as:

Principle 5.6. *Elga's Absurd Self-locating Principle of Indifference: Subjectively indistinguishable centred worlds should be assigned the same credence.*

Elga is saying that we should consider that indistinguishable mental states are equally probable, but *only* if they occur in the *same* uncentred world. The problem is that his justification for rejecting equiprobability across *different* uncentred worlds is unclear, beyond the mere fact that he thinks it "absurd". Not surprisingly, we will later see that this distinction is nearly impossible to carry over to an Everettian ontology, where the distinction between worlds is merely perspectival. Elga's distinction relies, on the other hand, on the idea that a *world* is a metaphysical absolute. For future reference, then, let's note a version of Elga's principle that *would* be consistent with an Everettian multiverse:

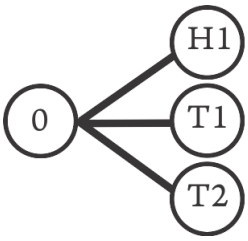


Figure 5.1: The Thirder Probability Tree: three subjectively indistinguishable mental states

Principle 5.7. *Everettian Self-locating Principle of Indifference: Subjectively indistinguishable, centred world branches should be assigned the same credence.*

Perhaps Elga would count this as effectively the same as his absurd principle-that-he-does-not-support, but there is no way to be sure without asking Elga himself, as his paper does not address Everettian ontologies.

Lewis summarizes the Thirder argument as follows:

1. After being told it is Monday, Beauty's credence is
$$p_+(H) = \frac{p(H1)}{p(H1) + p(T1)} = \frac{1}{2} \tag{5.6}$$
2. Therefore
$$p(H1) = p(T1) \tag{5.7}$$
3. And since
$$p(T1) = p(T2) \tag{5.8}$$
4. It follows that
$$p(H1) = p(T1) = p(T2) = p(H) = \frac{1}{3} \tag{5.9}$$
5. Note also:
$$p_+(H) = p(H) + \frac{1}{6} = \frac{1}{2} \tag{5.10}$$

Lewis notes that the change from p_- to p was produced by centred evidence $H1 \vee T1 \vee T2 = \neg H2$, or none at all.

The anti-Thirder argument, of course, is that because Beauty has received no new information on Monday, her credence cannot possibly change from 1/2 to 1/3. The Thirder response is that, while it may not *seem* that Beauty has received any new information (because she has received no new *uncentred* information), she has indeed received new *centred* information: she knows that she is currently in one of $H1$, $T1$ or $T2$, whereas the night before she knew that she was in day 0.

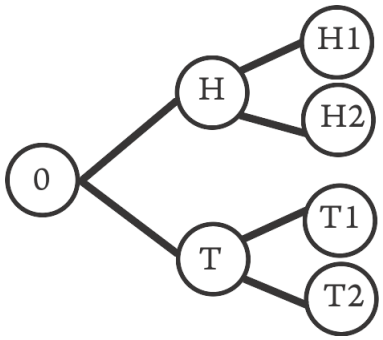


Figure 5.2: Thirder: Knowledge on Sunday

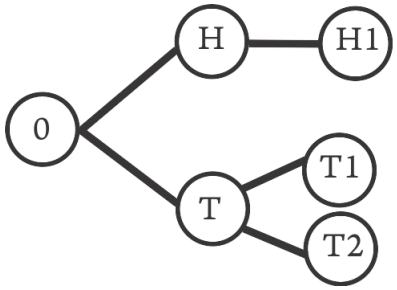


Figure 5.3: Thirder: Knowledge Changes on Monday Morning

The Halfer rebuttal is to argue that new centred "information" cannot actually be informative if it generates the epistemic state that was 100% expected already. Information lowers uncertainty in the receiver's mind, but how can uncertainty be lowered (or changed at all) if the person's epistemic state is exactly what she *knew* it would be all along?

The Thirder view, however, is that, because Beauty cannot trust her memory, she cannot trust the continuity of her experience. Hence, she must conclude that it could have been Tuesday (sitting at home) that she was experiencing right now, rather than any kind of experience in the lab at all. In effect, $H2$ effectively needs to be added as a possibility.

The fact that she finds herself at the lab at all *is* a kind of information, according to the Thirder. It is temporal self-location information, which is relevant information in this case, because it eliminates $H2$.

Remember: Beauty is not being asked the probability of a coin flip that is isolated from her. Her spatial-temporal information is affected by the flip, so information about her spatial and temporal situation validly influences her credences (or, so say the Thirders).

By this view, Beauty would in fact be justified to say the following to herself: "I cannot trust whether the time I am currently in is actually the next consecutive event after my subjectively

previous time; I could be either at LAB-Monday, or LAB-Tuesday. But I know I am not at HOME-Tuesday. These are the only three possible locations I could be in that affect the probabilities."

HOME-Tuesday affects the probabilities because if Beauty were there, she would know that the coin landed Heads, whereas HOME-Wednesday has no relevance, since if she were *there*, it would tell her nothing (since she would then be HOME-Wednesday no matter which way the coin flipped).

It is tempting to think that HOME-Tuesday shouldn't be considered a viable alternative to where Beauty finds herself, because she actually expected her next subjective experience to be exactly what it actually is. However, remember that Beauty cannot trust her memories here, so she cannot trust that there are no other alternative temporal-spatial locations for her, merely because she subjectively is in the situation she next expected herself to be in.

A way that this is often argued for by Thirders is to suggest an extension to the thought experiment, the so-called "million-mornings" version [23]. In this version, when the coin lands Tails, Beauty is awakened not *one* additional morning, but one *million* additional mornings, in the lab, with the same subjective experience. Now, when she wakes up and finds herself in the lab, it seems a bit more obvious that this is informative to her. There are a million mornings at home ruled out now, not just one. Surely she is more likely to find herself in one of those million mornings than in the single other morning that corresponds with Heads.

Elga suggests this result means that David Lewis was wrong in [127] when he said there was little difference in using the space of possible worlds and the space of possible centred worlds. To a Thirder, the relevant question one needs to ask oneself at any point is this: "is there any other experience I might have been having right now instead of this one?"

The answer on Monday morning is "Yes", in spite of the fact that this is subjectively Beauty's next consecutive experience, and there were no alternatives to it. The fact is, Beauty happens to *know* here that reality is not properly reflected by her sequences of subjective experiences. Since she knows that this is so (contrary to almost all usual experience), this means that she actually *can* receive new information by finding herself subjectively in *exactly* the very next subjective experience she expected to be in all along.

The defining characteristic of the Thirder position, from the viewpoint of probability interpretations, is that it assumes that subjectively indistinguishable states are equiprobable, and calculates the probability by counting subjectively indistinguishable states as the same (micro)state.

5.1.1.6 David Lewis and the Halfers

The Halfers take the following rule as a basic principle:

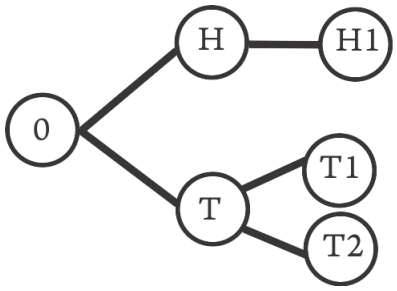


Figure 5.4: Halfer: Branches, not Leaves, determine probabilities

Principle 5.8. *No new information (even of the centred kind) can be received if one's epistemic state is exactly what one expected it to be.*

Lewis summarizes the Halfer position as follows:

1. By the above principle, new evidence is required for a change in credence.
2. $H1 \vee H2 \vee H3$ is not relevant information.
3. Therefore
$$p(H) = p(T) = \frac{1}{2} \tag{5.11}$$
4. Note also:
$$p_+(H) = \frac{2}{3} \tag{5.12}$$
$$p(T) = \frac{1}{3} \tag{5.13}$$

Lewis disagrees with the initial premise of Elga that $p_+(H) = 1/2$. The disagreement amounts entirely to the question of whether $H1 \vee T1 \vee T2$, *i.e.* whether $\neg H2$ is relevant to $p(H)$ and whether Elga can assume that $p_+(H) = 1/2$; in other words, that $p(H1|H1 \vee T1) = 1/2$. Elga says that it is, because once we know it is Monday, we can ignore Tuesday possibilities, and then there is no reason to prefer H or T.

But, says Lewis, when Beauty is told it is Monday, she is receiving (centred) information about the future (that she is not there). (And Elga agrees, since his credence measure changes, too.) Therefore we can't just assume that $p_+(H) = p_+(T)$.

Lewis in fact argues that
$$p_+(H) = p(H) + \frac{1}{6} = \frac{1}{2} + \frac{1}{6} = \frac{2}{3} \tag{5.14}$$

The fundamental feature of Halferism is that it does *not* regard indistinguishable mental states to be probabilistically indistinguishable, but instead regards metaphysical causation as the determiner of probability.

Figure 5.4 is Lewis's probability tree. This is a conventional probability tree, so that each branching is treated as additive[47]. The Halfer will consider these branchings almost as causative forces, so that there is some actual physical event that causes the branching to take place as it does. Of course, probabilities here are not literally ontic, but the coin flip as a 50-50 chance can be viewed as a coarse-grained model, and treated effectively as ontic for all intents and purposes. This is an example, then, of *epistemic chance.*

Nothing about these branches is defined in terms of subjective indistinguishability, although we are free to (optionally) group the resulting leaves of the tree according to such a criterion, if we wish. Consider the million mornings experiment, which is supposed to bolster the Thirder position. Surely, if there are a million mornings with one outcome and only *one* morning with the alternate outcome, it is the first outcome the will likely hold, reasons the Thirder. But the Halfer sees the situation differently. The Halfer points out that if it was a 50-50% coin flip that determined whether Beauty was set on a single-morning or million-morning course of events, then there are clearly even chances that she will be on either course. Without waking up with some *information* as to which course she is on, there is *still* a 50% chance she resides in the single-morning (not the million-morning) course of events.

Alternatively, we could imagine a million robotic copies of Beauty being made, while her original body is destroyed. The robots are awakened on the *same* morning in identical labs, having the same subjective experience. Now we are back to the same number of mornings as the original problem, but in one of them, we have created a million duplicates. It seems this should be the identical situation, probabilistically speaking. Again, from the Halfer's point of view, there is only a 50% chance that the million robotic copies were ever made in the first place, so without new genuine information—that actually reduces uncertainty—there is still only a 50% chance that Beauty is one of the copies, when she wakes up. Here, the rejection of Elga's indifference principle is very clear. We have a million and one identical copies of the same subjective mental state. Yet, a particular one of these has a probability equal to all the others combined.

5.1.1.7 The Generative Solution

To bring us back to the Everett debate: Elga is counting *observers*, and is, in fact, counting them according to something very much like (if not precisely the same as) synthetic unity.

[47] Perhaps confusingly, I already displayed the exact same tree (Fig. 5.3) in illustrating Elga's position. However, this same diagram is not a conventional probability tree for Elga, the way it is for Lewis. For Elga, the *leaves* of the tree are grouped according to subjective indistinguishability, and then a principle of indifference is defined on these groups, so that the branching of the tree does *not* obey additivity. For Lewis, the leaves are not necessarily equiprobable, but follow the rule of additivity at each branch point, leading back to a probability of unity at the root (making it a traditional probability tree).

The Halfer is counting the *branches* of the tree, and imposing an additivity requirement on subsequent branching. This sounds a lot like branch-counting in the Everettian debate.

The reader may therefore find it curious that, while I reject Everettian branch counting and accept synthetic unity, I am a Halfer, not a Thirder. A full treatment of the application to Everett will wait until a later chapter, but for now I will briefly state my reasons. While it may superficially seem that Lewis is counting branches, his "branches" are not analogous to Everettian branches. They are branches defined according to a metaphysical generative principle—*i.e.,* it is the act of flipping the coin that generates the two possible branches, and so these branches have equal probabilities, reflecting *not* the equality of branches, but the fairness of the coin (the generating condition). Use an unfair coin, and the branches will not be equiprobable. Subsequent coin-flipping events may (or may not) further bifurcate these branches. All such branch bifurcation, in Lewis's scheme, will be due to a physical, generative mechanism that creates multiple possibilities. This defines what one counts—one's indifference principle. Whether the states are subjectively indistinguishable does not come into play: although it may be allowed as the *category* we are finding the probability *of*, it is not an ontic countable. For Elga, indistinguishable states really *are* the same *micro*-state. For Lewis—at least on my generative reading of him—the *micro*-states can be generated by coin flipping (or some other generative mechanism) but they are not defined by their subjective features. Only *macro*-states are to be defined in that way. This means that the Halfer will never care that there are three subjectively indistinguishable states when determining his denominator count; the countables will be the generated branches in Fig. 5.4, not subjectively indistinguishable leaves.

Both Elga and Lewis need to delve into metaphysics to justify their positions, even though we are dealing clearly with *epistemic* probability, not ontic. In fact, when one looks at the Sleeping Beauty literature, it seems impossible to settle the Sleeping Beauty question without adopting some kind of metaphysical stance about what to count, and we have seen that Thirders and Halfers differ greatly in what they think we should count.

The Thirders' counting of subjectively indistinguishable experiences raises the question of justification, which Thirders seem short on. Elga seems to exclude possible worlds, and only counts indistinguishable events within one world. This is, first of all, problematic for application to an Everettian reality, where it is likely that either a huge number, or even all, possible worlds are in the wavefunction to some amplitude level. However, Thirderism is problematic for a more general reason, if we are looking for a source of objective probabilities: the *individual* entities it counts are defined subjectively. And they really *are* defined subjectively, not merely *a priori* synthetically.

I have rather glossed Elga's method for determining what to count—it is really only partially subjective, being more a combination of metaphysical considerations and subjective indistinguishability,

as it seems he chooses to count objectively distinguishable space-time locations as his outcomes, *then* groups them according to the question posed:

$$\begin{aligned} \varepsilon_H &= \{H1\} \\ \varepsilon_T &= \{T1, T2\} \end{aligned} \tag{5.15}$$

That is why he counts three microstates—if he really only used subjective indistinguishability to determine what to count, he wouldn't be able to get past a count of one. But is this really what he wants to count, in general? What if Beauty were to learn that, on Tuesday, if she is at the lab, there will be another coin flip, but the result will be ignored? There are now four objectively distinguishable outcomes. But would anyone suggest that we partition as follows?

$$\begin{aligned} \varepsilon_H &= \{H1\} \\ \varepsilon_T &= \{T1, TH2, TT2\} \end{aligned} \tag{5.16}$$

This would mean that $p(H)$ has changed from $1/3$ to $1/4$, just because we know that something happened that had no effect on the experiment at all. Yet, if we claim that we are counting only those things that "matter to the experiment", then why not as follows?

$$\begin{aligned} \varepsilon_H &= \{1\} \\ \varepsilon_T &= \{1, T2\} \end{aligned} \tag{5.17}$$

But, now, our events are not independent of each other.

Clearly, the intent is only to divide "1" (Monday) into as many distinguishable situations as matter to the question posed. Perhaps what Elga has really done is to first decide on what metaphysical requirement matters—in this case, he chose space-time locations, giving an initial sample space:

$$\Omega = \{1, 2\} \tag{5.18}$$

and then to further bifurcate the sample space according to the events in the question posed:

$$\Omega = \{H1, T1, 2\} \tag{5.19}$$

This could be perfectly valid, since one indeed might need to choose a more finely resolved sample space, if one is asked a more detailed question. However, the problem of indifference rears its head here: why could we not start with a completely different outcome set? For instance:

$$\Omega = \{H, T\} \tag{5.20}$$

In this case, the question posed does not require a finer resolution sample space, since the question is simply "H or T?" If the question is asked about the second coin flip, then we *would* need such finer resolution. Yet, this sample space yields the Halfer solution.

Does Elga's principle of indifference help us choose between these sample spaces? It is not clear that it does, since H and T are not really the kinds of things that are "subjectively indistinguishable", since they are metaphysical facts about the situation, not experiential characteristics. We could answer in the affirmative, if we interpret the principle as telling us that we can only count *experiences*—the kind of things that are or are not "subjectively indistinguishable". But then, as already mentioned, we should really only be counting one thing:

$$\Omega \quad = \quad \{HT12\} \tag{5.21}$$

since, *experientially*, there is only one subjective experience. Elga's problem is in insisting that metaphysics be used to generate more than one subjectively indistinguishable event, but then insisting that metaphysics has no role in determining whether these events are equiprobable or not.

At the end of the day, Elga cannot justify his choice of sample space. But Lewis can! Lewis chooses the coin flip outcome as the criterion for the sample space, because this is the distinguishing *metaphysical* (not experiential) generating condition or *cause* that determined the different outcomes. It was the coin flip that physically caused there to be more than one possible world under consideration—it "generated" the possible worlds (not literally, like an Everettian wavefunction, but possibilistically). Therefore, it must be the basis of the sample space. Elga's sample space is based on experiences, completely independently of their generating conditions (but with some non-generative metaphysical bias added to give preference to spacetime locations).

To bring things briefly back to the MWI: despite initial appearances, it is Lewis's position that is consistent with algorithmic synthetic unity, which, while defining *macrostates* in terms of synthetic unity, defines *microstates* in terms of generative conditions (in *algorithmic* synthetic unity, these generative conditions will be abstract computer programs). It is Elga's position that corresponds to an observer-counting synthetic unity (which I reject), which, like Elga, groups states *first* into synthetic (Elga would say subjective) indistinguishables, and then defines these as the *microstates*. Hence, Elga's (and the observer-counter's) *microstates* are Lewis's (and the computationalist's) *macrostates*.

Under the generative probability scheme, we cannot count synthetic *a priori* entities. If our resulting probabilities are to be objective—and be true for bunny rabbits as much as for people—then we must count something that really exists, not something that is perspectival. Synthetic unity does not justify perspectival ontic entities (microstates), in spite of its perspectival nature. Its use is restricted to preferencing one particular partitioning of these microstates into macrostates as

having *a priori* validity, since this partitioning is in terms of the unity of the observer's conscious states, and we are asking, in the first place, about the probability of one of this observer's possible conscious states being actualized. To leap from this (legitimate) use of synthetic criteria (to answer an already-synthetic question) to the counting of synthetic objects as our primary ontic entities, would be to imbue reality with a fundamental syntheticity, and to, in fact, succumb to the very anthropocentrism that synthetic *a priori* methods are so often (illegitimately) accused of.

Just as always, when working with objective chances, a subjective partitioning of our countables does not detract from the necessity that the countable itself be objective.

5.2 Self-selection and the Anthropic Principle

5.2.1 The Anthropic Principle

Just as Elga's principle of indifference addresses the problem of self-*location*, the anthropic principle addresses the closely related problem of self-*selection*. Imagine someone asks you the following question:

> *Why does the Earth have a (20%) oxygen atmosphere?*

Let's look at several possible answers one could give to this question:

1. The planet has plants, which give off oxygen.
2. The such-and-such initial conditions in the early formation of the solar system eventually lead to oxygen production.
3. So that we can breath.
4. Because, otherwise, we wouldn't be here to ask the question.

The first two answers are "causative" responses, and give the proximate reasons for oxygen being produced and maintained in our atmosphere (the second answer reaches back further in time to find the causative forces at work, but is ultimately the same sort of explanation as the first answer). These answers tell us *how* the oxygen got here. These answers are not wrong, depending on what was meant by the question.

The last two answers we might call "anthro-", or "human-based" answers, since they appeal to our presence to explain *why* conditions were such as to produce oxygen, in the first place. These answers are not wrong, either, depending on what was meant by the question.

Answer #3 is anthropocentric, since it explains the reasons as if they are purposive, *for* us. Such an answer may not be wrong, depending on the intent of the question, but is almost certainly not a *scientific* answer.

Answer #4 is an anthropic, or self-selective, answer. It may sound anthropocentric to the uneducated ear, but it is not. It explains the presence of oxygen by appealing to our presence, which could not have been otherwise, since we then would not be here to pose the question.

So let's replace the one original question with four more refined questions, that can legitimately be answered in these four ways:

1. What is causing this oxygen to be produced?
2. What were the original conditions that lead to this oxygen production?
3. Why should the Earth produce oxygen?
4. Why does the Earth have so much oxygen, when this is an extremely rare and unlikely condition for a planet?

Question #3 is normative, and not scientific in nature. These kinds of statements are sometimes put forth as examples of the anthropic principle, but I consider that misleading, and will not adopt that usage. Question #4 addresses not the *cause* of the situation we seek to explain, but its *unlikeliness*. The anthropic principle "explains" this unlikeliness, not by showing that the causes of the oxygen production are highly probable, after all, but by showing that the *result* of oxygen production is not only highly probable, but perhaps even necessary, *from our point of view*.

Definition 5.9. The "anthropic principle" or "principle of self-selection" states that any *a posteriori* fact must necessarily be consistent with our presence as observers capable of making the required observations necessary to establish the fact in the first place.

While this certainly "explains" the presence of oxygen, in some sense, it is a matter of debate whether this really counts as "scientific" explanatory power. Some would argue that, while not wrong, such a principle is no more "scientific" than the normative "should" question #3. A common criticism is that such a principle is a tautology, and therefore unfalsifiable and unscientific. Assuming (to oversimplify the situation) that oxygen is required for life, it goes without saying that our planet must have oxygen simply "because we are here". It would not be possible to falsify this hypothesis, since without the oxygen, we wouldn't be around to do the falsifying. However, this is no more scientific than "2+2=4"... it is simply analytically true, by definition.

This ignores the fact, of course, that much of science makes use of tautological principles. They are not testable, but that does not mean they cannot play a role in science (just so long as that role is not to be an hypothesis or theory). At the very least, it seems that adoption in science of the anthropic principle as a handy reminder to prevent one from making analytic mistakes would be hard to deny. Let's say there is an Earth geobiologist who is genuinely puzzled as to why the Earth has so much oxygen. His models and calculations, which are beyond reproach, clearly show

that such a possibility has a vanishingly low probability. He considers it a grand and unsolved scientific mystery. But only because he is forgetting the anthropic principle. We remind him of the principle; he realizes that his own necessary presence is enforcing a selection effect that eliminates all the non-oxygen producing planets from his sample space, and then all is well. His model of the geological and biological forces at work has not changed, nor been affected in any way by his being reminded of the anthropic principle. Nor did anyone ever consider this principle a theory or hypothesis about oxygen production. It "explains" oxygen production only by preventing us from making logical errors of reasoning.

Things become stickier, however, when the domain of the principle is expanded from "all the worlds in the universe" (as in the oxygen example) to "all the worlds in the multiverse" or even "all possible worlds". When the domain is restricted to the *observable* universe, it is used to "explain" unlikely-seeming events *within* our space and time, and we call it the "weak anthropic principle". Thus, it is easier to clearly see that the principle is not causative, but merely a "handy reminder". However, when the domain is bigger than what we can even in principle observe, it becomes a "handy reminder" of why the entire universe is like it is, and has the laws it has, in the first place. It is much harder to think of this as "just" a handy reminder, since it is actually selecting the type of universe we can be in—this means selecting out not just our *location* within the universe, but the very laws of the universe itself. Hence, at this level, the principle seems to be central to theory formation, even if not a falsifiable theory itself. Even if it cannot be part of testability—although I am not making that claim here—it seems that it could still be part of what makes one theory simpler than another. When the anthropic principle appeals to a domain larger than our observable universe, we call it the "strong anthropic principle".[48]

5.2.2 Pure and Empirical Anthropic Probabilities

There are two kinds of applications of probabilities in self-selective situations, which I will call *pure* and *empirical* (the latter is, strictly speaking, just a narrower application of the former).

Definition 5.10. A "pure" self-selection is a random pick from all possible conscious states (or from the entities that generate such states). A pure anthropic or self-selective probability, therefore, is the probability of obtaining a particular conscious state from such a pick.

Definition 5.11. An "empirical" self-selection is a random pick from all possible continuers of a given conscious state (or from the entities that generate such continuers). An empirical anthropic

[48] There are broader definitions than mine for both weak and strong principles, but these are the versions that I feel have scientific relevance, so I am restricting our attention to them.

or self-selective probability, therefore, is the probability of a continuer state, from the perspective of the continued state.

5.2.3 Synthetic-Unitary Self-selection

In my opinion, the strong anthropic principle—at least when considered as a guiding principle for quantum foundations—really only makes sense in light of the Everett interpretation and algorithmic synthetic unity. Without the Everettian multiverse, there is no way to make self-selection work on the scale of the observable universe. Without synthetic unity, there is no rational basis, that respects psychophysical parallelism, for a preferred basis and hence for the perspectival partitioning of the wavefunction into worlds. But if we try to use synthetic unity, without algorithmic information theory, we are left without an objective foundation for quantum probabilities, with no ontic entities to count, forcing us into either subjectivism, or something illegitimate like branch counting, which will push the anthropic principle into unfalsifiability (we will see why later in §5.2.4).

The following thought experiment will illustrate the use and rationale behind the principle of algorithmic synthetic unity. It is not about quantum mechanics (at least not directly), and is certainly not meant as a derivation of quantum theory (although it may be a step in that direction, a direction I will continue along later, but not here). Rather, this example has been intentionally chosen to lead to different consequences than quantum theory, so the reader can clearly see that *(1)* adopting the principle of algorithmic synthetic unity is independent of the context of quantum mechanics, but that *(2)* it can lead to a similar principle of superposition, even in a completely classical context.

Imagine a house with four equal and identical rooms, each of which has a door exiting towards a different compass direction. So there is a North room, and if you open the door there you will see a North-facing landscape—let's say the mountains. If you open the door in the South room, you see a forest, whereas the East door exits on a meadow and the West door opens to a view of the ocean. We will assume this house exists in a universe that obeys some sort of classical or quasi-classical physical laws—no nonlocal connectedness and no quantum weirdness is allowed here!

Now imagine we have four identical quadruplets placed in the four rooms at the beginning of their lives. For all practical purposes, they are in the identical physical state at this point, right down to the precise chemical makeup of each cell in their brains and the rest of their bodies. Likewise, the rooms they inhabit come equipped with everything they need to live, and are likewise each in an identical state, and—like Schrödinger's cat box—they are self-contained and 100% closed off from the environment.

Years later, the quadruplets have grown. All four spend their days doing exactly the same thing,

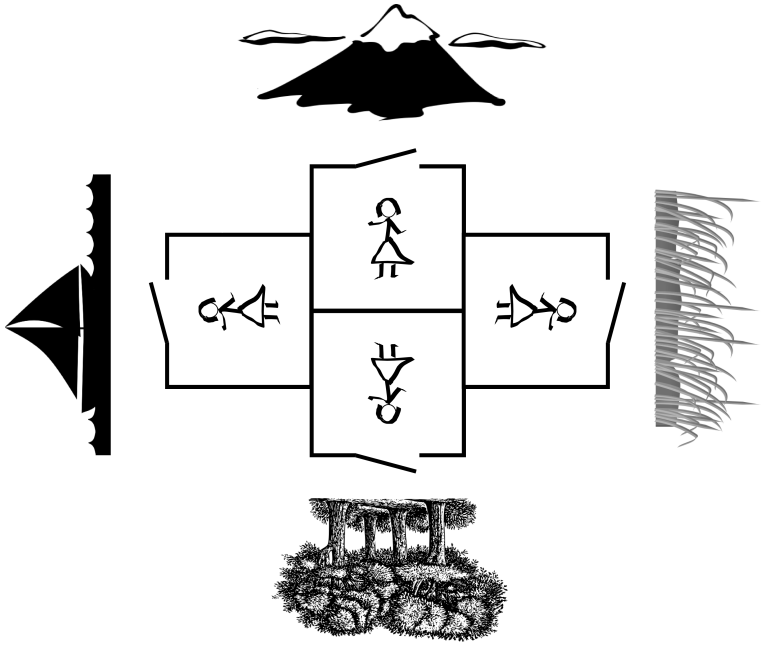


Figure 5.5: The Four Rooms

eating the same breakfast, at exactly the same time, and putting the plates away in exactly the same spot in the cupboard. They act in perfect synchrony, not because there is any communication between them, but simply because they are identical deterministic systems that started out in the same initial state.

Of course, if you believe in the Strong AI postulate—that an appropriately programmed computer could be fully conscious in the way human beings are, independent of its particular material substrate—then you will not mind replacing the quadruplets with four identically prepared and programmed robots. This will make the practical difficulties of isolating the rooms from the environment and putting them in the same effective initial state significantly less daunting, but it is not an essential part of the thought experiment.

The first philosophical question we have to answer about these quadruplet humans or robots (conscious entities, at any rate) is whether we are actually dealing here with one consciousness or four. You may want to think about that for a few moments, and decide what you think the correct answer is, before reading on.[49] In my experience, most people are fairly confident in their answer to this question. I will not argue here for what is the *right* answer. I will, however, argue that *if* you believe in Strong AI or servomechanism equivalence, then you *must* believe that there is, for all

[49] I need to thank Nick Bostrom here for bringing to my attention the benefit of asking this question before proceeding with *any* discussion that presumes a computationalist view of consciousness.

intents and purposes, only one consciousness in the house, not four. After all, the four people in the four separate rooms are computationally identical, and differ only in the actual material substrate their consciousness is instantiated in. So if, as Strong AI contends, the nature of personal identity and consciousness is completely substrate-neutral, it follows that once we accept Strong AI, we necessarily must contend that there is only one person in the house. I would be interested if there were anyone who purports to believe in the Strong AI thesis who might disagree with this, but it seems to me that such a position would be self-contradictory.

Imagine what it would mean to argue otherwise. A commonly repeated thought experiment meant to emphasize the consequences of Strong AI involves the idea of teleportation. What if we could somehow scan the state of your body and your brain into a computer?—reading off all the information of what each molecule is doing. Could we not scan this information, transmit it to some far-off place at the speed of light, then destroy your original body whilst generating a duplicate from the information at the faraway location? To most advocates of Strong AI, this is simply a high-tech form of transportation. To many others, it creates a brand new life, whilst murdering another. But so long as one takes the view that it is a form of transportation—since the transported person is a continuer of the original consciousness—I see no rational way to deny that two simultaneously running conscious programs with identical states are, in fact, the *same* consciousness. Why would a teleported identical copy be the same person, while a concurrently running identical copy is somehow not? The only possible distinction between the two cases would be that in the former, one copy is destroyed. But to argue that this distinction is meaningful, one would have to argue why destruction of one copy should have any affect on the personal identity of the other.

I will assume Strong AI for the four-room thought experiment, and I will assume also that this implies that identical copies of minds do not constitute separate consciousnesses.[50] The reader should be forewarned that this thought experiment makes three highly unrealistic assumptions:

1. that there are four similar black boxes, into and out of which no information can flow,
2. that we could somehow set up all four rooms, complete with a conscious being, in exactly the same initial state, and
3. that the universe in which this takes place operates quasi-classically and deterministically.

[50] It is possible that some might feel this thought experiment actually constitutes a *refutation* of Strong AI, given that we end up with four *obviously* separate individuals, and only one consciousness. However, we need to resist here what I call the "Chinese Room Fallacy", in which one sets up a thought experiment which is possible in some highly theoretical sense, but not in any practical sense—usually by imagining that something immensely complex can be treated as if it were very simple—and then tries to argue a point based solely on one's intuitive gut reaction to the (in practice) absurd result of the thought experiment. Searle did this in his Chinese Room experiment, when he asked us to imagine a person simulating an entire brain of another person algorithmically—by hand with paper and pencil—and then drew on our intuitive gut reaction that there could not really be "another mind" in the room, to conclude that minds cannot be algorithmic in nature.

These assumptions are not pathological, so long as we do not take their consequences too seriously. This is not a real-world scenario, but a hypothetical-world scenario intended to help us think more clearly about our own world. Note that there is no denial here of *metaphysical materialism* (the idea that the universe is ultimately made up of matter and energy)—in fact, we are essentially assuming materialism, given our assumption of quasi-classicality. Rather, there is only a denial here of *materialism with respect to conscious identity* (the idea that one's personal identity and continuity of consciousness depends on the particular material substrate in which it is instantiated). An advocate of the Strong AI postulate need not be anti-materialist, metaphysically, so long as there is no necessary connection drawn between conscious continuers and particular material substrates. (Indeed, many Strong AI advocates *are* materialists.)

Given this unrealistic scenario, and given our philosophical assumptions, we can conclude that there is only one consciousness in the house—and thus only one person—even though there are five separate biological bodies and brains. So let's name this person, this consciousness, "Liz" (no need to assign four names, as this is just one person).

Returning now to the four rooms, it seems that every so often the Liz's decide to try the door (or should we say doors?). Until now, the door has always been locked, and so this one room (rooms?) constituted Liz's entire observable world. Little does Liz know, that the door(s) were set up with a time-lock mechanism, scheduled to auto-unlock on her thirtieth birthday. So, inevitably, when Liz turns thirty, the doors unlock. Liz is sitting at the kitchen table, at the time, and actually hears the noise of the lock mechanism going kir-CHUNK! as the door unlocks. Curiosity aroused, Liz approaches, turns the handle, and opens the door.

What does she see? The ocean? A meadow? A forest? Mountains? Or does she see some weird "superposition" of these things? Remember, this is not a quantum universe. There is no superposition principle built into its physical laws. It is a classical—or at least quasi-classical—deterministic universe, following laws that obey local causality.

The answer, I think, is clear. While there may be only one person, there *are* four physically separate rooms. The four Liz's constitute a single person only because, until now, they were running the identical computational processes. The moment Liz opens the door, there are now four *different* Liz's, seeing four different scenes. Liz has divided from one person into four.

There is, of course, no reason to suppose some mysterious "splitting mechanism" here. Liz did not really "split" and neither did her world. The so-called split is just a perspectival phenomenon. Rather, Liz, by opening the door, allowed information to flow from the environment into the previously-black box of her mind. She *gained information* as to which side of the house she was actually on. In a sense, she had been there all along—at least she can say this *after* the split. But before the split,

she could not possibly claim to be in only one of the rooms (without contradicting Strong AI) since the mental and conscious states in all four rooms were identical. If we really take seriously the idea that Liz, as a person, as a consciousness, is defined by her mental computational processes, and *not* by her material substrate—to the extent that we would accept an informationally teleported version of her as really *her*—then before she opens the door, while she is still sitting at the kitchen table, we really have no choice but to say that, beyond the door, exists a superposition of scenes, which collapses down to one scene when she opens the door.

Now—just for fun—suppose we invert the handedness of just two of the rooms, immediately after putting the baby girls into the rooms. Thus two of the Liz's, at the beginning of their lives, are right-handed, and two are left-handed. This might be a difficult job of molecular nanotechnology, but it is in principle doable by mundane atom-shuffling. If we make Liz an artificial conscious robot, the job becomes much easier, as we can simply construct two of the robots and rooms by a right-handed plan, and the other two by a left-handed plan. Either way, the result is that the four rooms (and the girls) remain *internally* identical. The handedness inversion makes absolutely no difference whatsoever in what any of the Liz's see or experience in their rooms.

So now we have a situation where *one* person, Liz, for the first thirty years of her life, is living with a superposition of completely different landscapes just outside her door, and is *herself* in a superposition of being both left and right handed. The result of the handedness superposition is that, when the four—now different—Liz's walk out of the house to explore the world around them, when they meet the other three Liz's, each will discover that one of the others is exactly like them, but that the other two have opposite handedness.

Now let's move on to the question of probabilities. What if we were to ask for the probability of Liz's seeing mountains when she opens the door? Since Liz has no conception of the actual situation she is in, this is not a subjective probability. It is objective, since it is what it is, even if Liz has no idea that anything that happened to her was anything but deterministic. (This thought experiment thus serves as a proof that the notion of nontrivial single-case objective chance at least makes conceptual sense.)

It would seem rather uncontroversial that the probability is $1/4$, and this is based on the assumption that the four rooms are equally probable. This all seems straightforward, given the setup for the thought experiment—what reason could there be to prefer one room over another, probabilistically?

However, let us modify the experiment slightly. Instead of having the doors open out to the external world, they each open instead to a small, windowless foyer. In the North foyer is a small table with a red rose in a vase. The door leading outside is closed. Each of the other three foyers is identical in every way, except that the vase has a white lily in it, instead of a red rose.

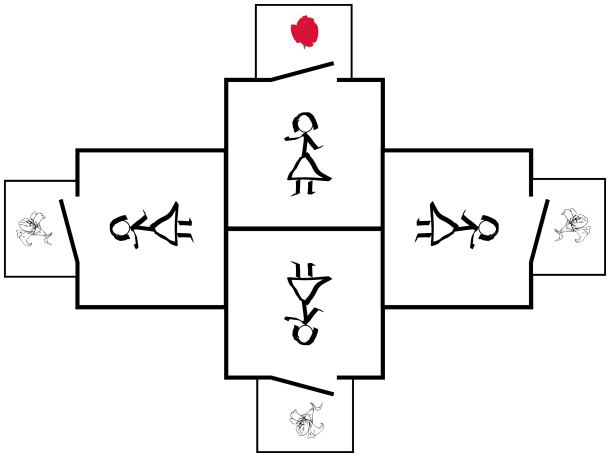


Figure 5.6: The Four Foyers

Now we ask, when Liz opens the door, what is the probability that she will see a lily? Clearly, the probability is 3/4. Yet, if we consider that, from Liz's perspective, she has essentially split into *two* copies of herself—not four—one of which sees a rose and the other a lily, we might wonder if the probability should be 1/2. Yet, if Liz were able to repeat this experiment over and over again, she would clearly observe a lily three quarters of the time.[51]

Assume now that we tell Liz that her statistically-observed "probabilities" are objective, and her universe deterministic. Ignore how we manage to convince her of this, and assume that she simply believes what we tell her (and, of course, we are not lying; these facts are true). Knowing that her observed probabilities are not due simply to her ignorance, and yet are generated deterministically, she might feel that she has no choice but to theorize that the other possible results of her observations are "equally real", and that there exist equally real other Liz's out there, experiencing the other outcomes. She might then adopt the perspective that every time she makes an observation, her world essentially "splits" in two. She would not know if this was a real splitting mechanism, or simply her perspective. But either way, she might wonder why it is that her probabilities do not align with the "world" or "branch" count, which would seem to demand a result of 1/2. The rational conclusion for Liz to draw is surely that the "worlds" or "branches" are merely perspectival, and that the empirical probabilities she measures are what would result if she could actually count the real countable entities of her universe. She does not know what these ontic entities are, so she just gives them some random technical name. Thus does Liz, through repeated empirical observation, come to a theoretical understanding of the objective existence of the other rooms.

[51] The experiment can easily be modified to allow these repeated trials by simply increasing the total number of rooms, and making each foyer's outer door open up to yet another foyer, which opens up to yet another foyer, and so on, to some large number of iterations that Liz will never be able to exhaust in her lifetime.

Of course, we had to *tell* Liz that her probabilities were objective. Without this additional information, she probably would simply assume that her probabilities were epistemic, and that she was living in a house with one large room and a single long chain of foyers, three-quarters of which have lilies and the rest roses. While synthetic unity demands a superposition of states for Liz, which evokes ideas from quantum mechanics, we do not find here the kind of interference effects that we see in quantum mechanics, that would allow Liz to infer the existence of the other rooms on her own, purely from her empirical observations. (And I will argue that it is just such interference effects that really create the biggest problem for amplitude counting in Everettianism.) Liz's probabilities can be calculated with simple, well-behaved sums, without destructive interference—by counting the ontic entities of her universe (the rooms that her copies occupy). Each room is *one* classical possibility to be added to the count, and each room with a white lily *adds* to the total chance of Liz's seeing a white lily.

But quantum amplitudes interfere. If they were really like the counts of Liz's rooms, we would not expect multiple possible contributions to the *same* outcome to destructively interfere with each other. To accept amplitude counting—on the kind of *a priori* grounds the Born rule objectors are insisting on—we need an *a priori* explanation for this interference. This will be the main goal of Ch. 6, where I will show that adopting an *algorithmic* view of probabilities means that some kind of interference effects are to be expected. We do not see them in Liz's world, because we assumed a materialist (and hence, by definition, non-algorithmic) ontology for it.

While the four rooms experiment, as is, may not give a rationale for interference effects, it does give us reason to see many of the features of quantum theory as non-mysterious, and even expected in a deterministic universe containing multiple copies of the same consciousness. There are two important take-away messages here (assuming that we accept Strong AI):

1. A perspectival superposition of states, resulting in an effective collapse during the act of observation, can be the natural consequence of a deterministic, classical universe, appropriately prepared.
2. There is no reason to assume that the objective probabilities in such a universe will be in accord with a count of branches (or worlds, or observers, or outcomes). In fact, it would be mere coincidence if this were the case.

Keep in mind, again, that this is a completely quasi-classical universe with local causality. Yet, by adding the postulate of Strong AI, and creating identical synchronized copies of our observer[52], we find ourselves forced to posit the idea of superpositions and instantaneous, nonlocal collapses.

[52]In fact, the copies do not really need even to be synchronized. Just as Liz's spatial location is a synthetic matter, so is her temporal location (if Strong AI is literally correct). Assume we set up the original four rooms so that the North Liz lives at a time in the remote past compared to the other four, and the South Liz operates at twice the speed of the other Liz's. There is still a superposition of four states when Liz opens the door, since all four copies are experiencing the same consciousness at that moment in Liz's time (the absolute time clock of her universe does not,

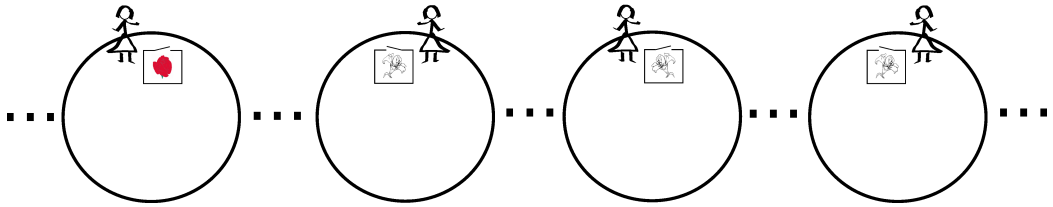


Figure 5.7: The Four Planets in Galaxies Far Far Away

While the four rooms thought experiment seems completely contrived, it is feasible that the actual universe could be like this, if it turned out that the universe was infinite (or at least, "big enough"). Imagine, instead of four rooms, four planets in galaxies vastly remote from each other in space and time, in a deterministic universe that is infinite, or at least big enough to allow that there just happen to be, by sheer coincidence, four nearly identical copies of the same planet, with copies of Liz and the flowers, just as in the house with the foyers—except that now the flowers are enclosed, like Schrödinger's cat, in a box, which Liz has to open in order to see what kind of flower she has.

It has been speculated that if the universe is infinite in extent, then one should expect to find copies and near-copies of whatever is physically possible somewhere within it, and that the anthropic principle will play a role in such a "Big World" cosmology similar to that which it plays in multiple universe models [22]. In such a model, if we accept Strong AI and its rational consequences (ASU, the informational teleportation of minds, *etc.*) then we automatically have a universe in which "superpositions" occur. One might argue that these superpositions are very different from quantum superpositions—after all, Liz's copies are in remote galaxies of the same three-dimensional space. This is nothing like Everettian superposition, where the copies can (but do not always) occupy the "same" 4-D space-time location (due to the fact that these locations are (more precisely) projections onto orthogonal 4-D subspaces of a larger-dimensioned Hilbert space).

But this is the wrong way to look at it! If synthetic unity is correct, then one's "4-D space-time location" in the quasi-classical universe is determined *by* synthetic-unitary considerations. The different Liz's in different galaxies would *not* be in remote locations, by definition, since "location" is a synthetic *a priori*, not analytic, property of things in her universe. The different near-identical copies are in the *same* space-time location, necessarily, since they contain the same consciousness, and space-time location is relative to the synthetic unity of consciousness. From the *a priori* synthetic

---

then, correspond in any direct way to *her* time). When she opens the door, she splits into four observers. One sees the North landscape, but meets no other Liz's when she explores the grounds of the property. The other three Liz's do meet each other, but one appears to be "speedy Liz", talking and walking and generally acting at twice the speed of the other two.

perspective, Liz's superpositions are now looking much more like quantum superpositions.

In any case, the main purpose of this thought experiment has not been to show that Liz's superpositions are *exactly* like quantum superpositions. They lack many of the relevant features, such as interference effects (although we will see later that an algorithmic rather than materialist ontology will, in fact, introduce such effects into the picture). But for now, the main point is that, even within a quasi-classical, deterministic metaphysics, macroscopic superpositions are the inevitable consequence of three not-entirely-implausible assumptions:

1. Strong AI,
2. Synthetic unity, and
3. A big (and varied) enough ontology.

Like Everettian superpositions, these synthetic classical superpositions are perspectival, and the corresponding "collapse" is, while non-local, clearly not a physical mechanism, and so could not be used to send instantaneous signals. Rather, the collapse occurs simply because what is "really" out there in Liz's environment depends on how much information about that environment is encoded in her brain—not because Liz's consciousness somehow has the spooky ability to reach out and influence its environment, but because the very idea of "environment" is *already* a perspectival, synthetic construct, and not a property of reality, as such. This is not a mentalistic conception of Liz's reality, however, where her environment is somehow created by her mind. Liz's reality is completely materialist down to its foundation, and in no way idealistic or mentalistic in nature. It is only *qua* Liz's environment that it is a synthetic construct—since viewing it *as* an environment depends on the artificial separating of Liz's purely material universe into "Liz" and "everything else".

5.2.4 Falsifiability of Anthropic Theories

There are numerous reasons the Anthropic Principle (or AP) is controversial. I shall attempt to respond only to the objections most relevant to this dissertation, and will therefore focus on the "Many-worlds Anthropic Principle" (MWAP), which is based on an ensemble of worlds (which may or may not be Everettian).

Definition 5.12. The "Weak Anthropic Principle" (WAP) is the observation that whatever features the universe has, these must be consistent with our presence (*i.e.*, the presence of conscious life).

While this principle may be useful, as it prevents us from forgetting about the rather obvious (but sometimes overlooked) empirical fact that we are here, it does not claim the greater explanatory power of the Many Worlds Anthropic Principle (MWAP), which claims to explain *why* the universe has these features, not merely remind us that it obviously does.

The MWAP is one version of the "Strong Anthropic Principle" (SAP), which includes any version of the AP in which the emergence of intelligent life is in some way "necessary". Carter's original distinction between weak and strong principles [41] places a wide variety of different ideas under the SAP umbrella, so much so that any general critique or defence of the SAP is probably doomed. The SAP label includes the MWAP, which seeks to explain the presence of conscious life by postulating a simple mechanism which may have given rise to it (self-selection given the existence of multiple worlds). However, the SAP is also sometimes taken to include the hypothesis of Intelligent Design, which is a completely different concept, which, if it did provide any explanatory power, would do so on a completely different basis (critique of this or similar ideas is outside the scope of this dissertation).

I will also not consider of interest objections to the AP based solely on the idea that it is teleological and/or anthropocentric, since such arguments clearly do not apply to the MWAP. I will also not consider here objections against the MWAP that attack only its "many-worlds" component, as such objections are broader than the topic of the AP, and are dealt with in other sections of this dissertation. In this section, we are focussing on specifically defending the anthropic principle within the context of many-worlds.

The most serious critiques against the MWAP, such as that of Smolin [202], seem to be centered on the idea that it violates basic principles of sound science, in particular Popper's [158] principle of unfalsifiability, and that it thus cannot be a scientific idea. I will look at Smolin's argument in some detail, as it seems to me to be a logically sound argument (I will quibble with its axioms in this chapter, not its logic) which has been highly influential, and is also particularly relevant to the issues in this dissertation. Smolin critiques only the MWAP, which is more or less what he means by "the anthropic principle". Smolin's notion of many-worlds is by no means restricted to Everettian ideas, and indeed, Smolin is more influenced by AP variants that have come up within cosmology and string theory. In addition, Smolin is *not* in any way arguing against many-worlds theories—indeed, he has his own many-worlds hypothesis that he has put forward (the idea that universes actually reproduce, and undergo a selection process much like natural selection [201]). He is arguing only against the use of many worlds *in combination with* the AP.

My own initial response to the general claim that the AP is unfalsifiable, and therefore outside of science, is to refer back to Popper [158] and make note that not all propositions that are "a part of science" need to be falsifiable. Thus, we can reject the more extreme statements that jump immediately from "unfalsifiable" to "cannot be part of science". The principle of survival of the fittest, for instance, is a very important principle upon which the theory of evolution rests, but it is, in itself, essentially tautological: "what is best at surviving is (more or less) what survives." The

principle is important, nonetheless, to the formation of the theory of evolution, and its continued motivation, and to our preference for it over theories that may explain the data equally well, but lack its simplicity and elegance. Having said this, however, I cannot help but agree with Smolin that if the AP is indeed unfalsifiable—in the way that he thinks that it is—it would be in quite a spot of trouble, at least with respect to the role it is typically imagined to have these days.

Popper identifies [158] at least four ways that statements that are scientific may be legitimately unfalsifiable. These are statements that have to do with:

1. *Theory consistency and coherence*: statements that help eliminate theories due to internal inconsistency or incoherence.
2. *Theory simplicity*: statements that help eliminate theories due to unnecessary complexity.
3. *Theory falsifiability*: statements that help eliminate theories due to unfalsifiability.
4. *Theory formation*: statements that play a role in the creative process of developing and proposing new theories.

Popper specifically disclaims the idea that this list is supposed to be exhaustive, so even here, he is not trying to put blinkers on science and eliminate new methodologies *a priori.* However, his list is certainly a good starting point, and seems to cover the basic uses in science of unfalsifiable statements. (It is worth noting that Smolin's argument against the AP is, itself, an instance of #3, and as such—if it is a valid argument—is unfalsifiable.) I do not, however, believe that the use of the AP that Smolin is primarily concerned with falls exclusively under any combination of the above categories. The first three are primarily negative, and thus concerned with *eliminating* theories for *a priori* reasons. Most advocates of the MWAP, it seems to me, view the anthropic component of their theories as an essential part of why the theory has explanatory power, and how it generates empirical predictions. It is intended to be fully a part of the machinery of the theory itself, rather than being an argument *about* the theory, or about competing theories. Likewise, its function is not merely as a part of theory formation (#4 above). An unfalsifiable statement that serves *only* to help form a theory does not play any further role in the scientific process (at least, not until the theory is falsified, and in need of modification or replacement, in which case the theory formation stage of science comes into play again).

I believe, therefore, that Smolin is on firm ground to be concerned about the falsifiability of MWAP-based theories. The kinds of theories he is concerned with (which would include my own approach in this dissertation) are claiming that there is an anthropic selection effect at work, that is responsible for the laws of physics being what they are. In a sense, there is a "random pick" from a "bag-full" of some *a priori* entities (such as universes), and this selection effect allows us to predict what the observed laws of physics are (most probably) going to be. This "bag-full" subsumes

all possible universes, within the parameters of some theoretical framework (and Smolin recognizes that we may need some kind of prior metaphysical framework to get us started). However, beyond this prior framework, there is no actual physical mechanism that is also part of our theory that determines what kinds of entities are in the "bag" (if there *were* some such mechanism, then there may still be a selection effect of some sort, but it would not be *merely* an anthropic effect, which is by definition a *self*-selection effect—a random pick merely by virtue of our existence within the system as observers). Note that it is the assumption of self-selection that presumably allows our theory to make the predictions that it makes about what form the laws of physics will take.

Definition 5.13. We will define an "anthropic theory" as a theory that posits the existence of an ensemble of *a priori* entities, of which we, as observers, are a part. From the mere fact that we are a part of this *a priori* ensemble can be deduced laws of physics, and thus empirically testable (falsifiable) consequences.

If Smolin could show that the anthropic-ness of any such theory (the exclusively *a priori self*-selective nature of its selection effect) were unfalsifiable, then it would follow that *any* predictions at all that we wished to make about the observed laws of physics would be consistent with self-selection. This would mean that *all* theories were "anthropic" in the above sense, no matter what predictions they made about the laws of physics. But, if that were the case, it would follow straightforwardly that no such theory could actually make any such predictions in the first place, since the presence of self-selection would be irrelevant to the specific predictions the theory made (given that all theories would be consistent with self-selection).

This is, in fact, just the kind of unfalsifiability proof that Smolin attempts. If it were to carry through, I believe it *would* show that anthropic theories are unscientific. It would not *a priori* bar the invocation of the anthropic *principle* from all scientific discourse, but when the AP is invoked in theoretical physics, it tends to be in the context of anthropic *theories*, not merely in the context of using the anthropic principle as a negative criterion for eliminating theories, or as a creative aesthetic or general metaphysical framework. So, while Smolin's argument (if successful) would not prevent us from ever invoking the AP in a scientific context, it would severely limit its use, and would place anthropic *theories*, as defined above, outside the boundaries of science.

Smolin calls his *a priori* self-selective ensemble "$\mathcal{M}$", and defines it specifically as an ensemble of universes. In Smolin's words:

> ... the version of the anthropic principle that is usually put forward by its proponents as a scientific idea is based on two premises.
>
> **A** There exists (in the same sense that our chairs, tables and our universe exists) a very large ensemble of "universes", $\mathcal{M}$ which are completely or almost completely causally

> disjoint regions of spacetime, within which the parameters of the standard models of physics and cosmology differ. To the extent that they are causally disjoint, we have no ability to make observations in other universes than our own. The parameters of the standard models of particle physics and cosmology vary over the ensemble of universes.
>
> **B** The distribution of parameters in $\mathcal{M}$ is random (in some measure) and the parameters that govern our universe are rare.
>
> This is the form of the Anthropic Principle most invoked in discussions related to inflationary cosmology and string theory, and it is the one I will critique here.

The first thing to note about Smolin's argument is that he is actually critiquing a very specific kind of self-selective ensemble. The prior entities that make up his $\mathcal{M}$ are constrained in **A** above:

1. to be causally disjoint spacetime universes, and
2. to operate under the parameters of the "standard models of physics and cosmology".

Thus, his argument might be valid, while still not addressing other forms of the MWAP that do not meet his criteria for $\mathcal{M}$ laid out in **A** (I will, in fact, argue that this is the case). The criteria laid out in **B** is more or less the standard pre-requisite for "self-selection", applied to the ensemble defined in **A**.

Smolin's argument, in brief, is that, since we know our universe is part of the ensemble (presumably because we are here), we know that at least one member of the ensemble will always have any property we discover about the universe. So no matter what experiment we ever do, for all time, the result will always be consistent with **A** and **B**, no matter what it is, or how unlikely it might seem.

> Here is the basic argument why a theory based on **A** and **B** is not falsifiable. If it at all applies to nature, it follows that our universe is a member of the ensemble $\mathcal{M}$. Thus, we can assume that whatever properties our universe is known to have, or is discovered to have in the future, it remains true that there is at least one member of $\mathcal{M}$ that has those properties. Therefore, no experiment, present or future, could contradict **A** and **B**. Moreover, since, by **B**, we already assume that there are properties of our universe that are improbable in $\mathcal{M}$, it is impossible to make even a statistical prediction that, were it not borne out, would contradict **A** and **B**. [202]

For instance, let's say we have an anthropic version of the theory of gravity, that claims to explain gravity as the probable result of a random pick from the ensemble of worlds. We now test our theory of gravity by holding a rock out and letting go, to see if it falls to the Earth. The theory predicts that the rock will fall to the ground. Many attempts have been made to falsify this theory of gravity in the past, and it has survived, so it is a well corroborated theory. Now, what if we hold the rock out and let go, and find suddenly that it floats in mid-air. So long as we discount the possibility of other forces at work, or errors of observation on our part, the theory of gravity is falsified. Yet, the anthropic part of our theory necessarily survives, because, now that we know gravity can be violated in this fashion, we know we are in a universe that does not follow the law of gravity, and that such

a universe is a member of the ensemble, and so we have *not* violated the MWAP. Likewise, for *any* experimental result of *any* kind, we can say the same thing. Whatever aspects of our theory may be falsifiable, the anthropic principle will never be among them.

Smolin next modifies **B** into a template for theories that might actually be testable, while retaining the many-worlds feature. Smolin's intent seems to be to save premise **A** (the many worlds part), by modifying **B** (the anthropic part) so that it has an empirical or falsifiable component, yielding **B´**:

> **B´** It is possible nevertheless, to posit a mechanism, $\mathcal{X}$ by which the ensemble $\mathcal{M}$ was constructed, on the basis of which one can show that almost every universe [that contains conscious life] in $\mathcal{M}$ has a property $\mathcal{W}$, which has the following characteristics:
>
> > 1. $\mathcal{W}$ does not follow from any known law of nature or observation, so it is consistent with everything we know that $\mathcal{W}$ could be false in our universe.
> > 2. $\mathcal{P}$ There is a doable experiment that could show that $\mathcal{W}$ is not true in our universe.
>
> If these conditions are satisfied then an observation that $\mathcal{W}$ is false in our universe disproves **A** and **B´**. Since, by assumption, the experiment is doable, the theory based on these postulates is falsifiable. [202]

Smolin makes it clear that **B´** is not actually a particular postulate, but a class of postulates, depending on which construction mechanism $\mathcal{X}$ one chooses for $\mathcal{M}$, which will depend on the details of the specific theory being proposed.

I do have a few quibbles and qualifications about Smolin's **B´** that are not, in the end, substantive to the core of the argument, but are worth covering before we go any further, to avoid possible confusions.

1. *Almost every life-containing universe in the ensemble has property $\mathcal{W}$.* Smolin originally constrained the universes to contain "life", which I which change to "conscious life". After all, if the properties of universes that contain life turn out to be radically different, in general, from those containing consciousness, then it will be only the properties of those containing consciousness that are relevant. We cannot possibly exist in a life-containing universe that is inhospitable to *conscious* life. However, Smolin refers even to "life" only in a footnote, as an optional feature. But in fact, without it, the "anthropic principle" Smolin is referring to, will have no real philosophical motivation, nor explanatory power.
2. *Property $\mathcal{W}$ does not follow from known natural laws or observations.* Smolin is unnecessarily restricting the possible features here. It is really only necessary to state that the property does not follow "from any known observations." Natural laws are not "known", in the strictest sense, anyway—and they are still considered fully falsifiable, in a Popperian philosophy, even when they are extremely well corroborated. The law of gravity might be called a "known natural law" as well as anything, yet we can still imagine experiments that might falsify it, so it remains falsifiable and therefore scientifically testable. Because natural laws are of a universal character, they tend to have this feature, whereas observations are particular, and

so do not. Thus, all we really need say here is that *the property $\mathcal{W}$ is not mandated by any known observations.*

3. *There is some doable experiment that could actually contradict $\mathcal{W}$.* The experiment, of course, does not really have to be "doable" *right now.* Nor need we even be capable of giving a precise description of what the experiment would look like: a set of general characteristics of, or constraints on, the proposed experiment would be enough to establish that there is possibly a doable experiment that could be designed in detail at some point in the future, and the theory would remain "falsifiable" until such time as someone shows why there is no doable experiment that could ever meet our criteria. To require practical and immediate doability would pretty clearly be counterproductive to the progress of science.[53]

Smolin claims that **B´** is a *modification* to the anthropic principle—presumably, such that it is no longer an example of it. What **B** really says is that the ensemble of worlds is random, by some measure, and that we are a random (and unlikely) pick from this ensemble. **B´** modifies this so that: *there is a mechanism $\mathcal{X}$ by which the ensemble can be constructed, which has the above three characteristics.*

The point here is that the ensemble for **B´** is no longer random, having a specific physical mechanism responsible for its generation, a mechanism which presumably will produce some possible worlds, but not others—or at least produce some less frequently. In Smolin's words:

> This is produced by taking properties allowed to vary within the theory, and selecting their values randomly, according to some measure on the parameter space of the theory. By random we mean that the measure chosen is unbiased with respect to choice of hypothesis as to the physical mechanism that might have produced the ensemble. [202]

Smolin's states that for an MW theory to be falsifiable, there must be a *high* probability of a random pick from $\mathcal{M}_{\mathcal{L}}$ having property $\mathcal{W}$, and a *low* probability that a random pick from $\mathcal{R}_{\mathcal{L}}$ will have $\mathcal{W}$, which we will notate as follows:

$$\begin{aligned} p(\mathcal{W}\,|\mathcal{M}_{\mathcal{L}}) &\simeq 1 \\ p(\mathcal{W}\,|\mathcal{R}_{\mathcal{L}}) &\simeq 0 \end{aligned} \tag{5.22}$$

This is why his own theory—with universes that reproduce and evolve by natural selection—is *not,* according to Smolin, an example of the AP, since his ensemble is generated by an $\mathcal{X}$ that is in no way rigged to favour life or consciousness, or even the features of a universe that would be *a priori*

[53] I would suggest, therefore, that the onus is on the critic to show conclusively that a proposed theory *is* indeed unfalsifiable, before labelling it as such. On the other hand, until a falsification test is specified, the proposed theory is still just that—a proposal—and remains in the "theory formation" stage of science, and hence should not really be granted the full status of a "theory" until some kind of experiment has been at least vaguely described (the dividing line between these categories may, of course, be fuzzy). Neither, on the other hand, should a proposal be labelled as "unfalsifiable and therefore unscientific", without a firm proof of such, since such a labelling (which Smolin attempts to place on the AP) does far more than challenge the proponents of the proposal to design doable experiments—it actually attempts to cut off and halt any further investigation of the idea.

favourable to them (although they may *accidentally* be favourable to them, if for example, it just so happens that the features that favour reproduction of universes also favour life).

The crux of Smolin's argument is that MW theories that do *not* meet (5.22) cannot, even in principle, be falsified. In other words, our selection (random pick) from the universe ensemble cannot be from the set of all possible universes, or even all universes that are possible given the parameters allowed by our theory. Moreover, our falsifiable property must be present in much greater abundance in our non-random ensemble than in a random one. And, finally, the difference between the random and non-random ensemble must not depend *a priori* on life or consciousness.

The most unconstrained possible notion of randomness-within-our-theory would be the case where $\mathcal{M} = \mathcal{R}$. In that case, since $\mathcal{M} = \mathcal{R}$ and $\mathcal{M}_{\mathcal{L}} = \mathcal{R}_{\mathcal{L}}$, (5.22) will clearly not apply, regardless of our definition of randomness. For more constrained notions of randomness, one could possibly quibble with Smolin's invocation of random-within-a-theory: could one not simply choose to specify the parameters of the theory in such a way as to make an otherwise random ensemble non-random, or vice-versa? However, this is not a substantive objection. If the AP has any explanatory power at all, there must be *some* legitimate notion of random-within-the-theory, or else the whole idea of a *random* selection effect could not be formulated.

There is, however, still a problem with Smolin's general conclusions—which tend to be very sweeping—even though his argument is essentially sound. The problem is not with his logic, but with his assumptions, and where he goes with his conclusions. He assumes both **A** and **B**, showing that all resulting theories are unfalsifiable, and then argues for modifying **B**, yielding a *non*-anthropic theory. But, of course, we could also modify **A**, by choosing a different sort of ensemble, while retaining **B** (and hence the anthropic nature of our theory). Recall that **A** constrains the members of $\mathcal{M}$ to be causally disjoint regions of space-time ("worlds", or "universes"). However, these members must also be reasonable *a priori* entities, or else it makes no sense to talk about a random pick from them being a "self-selection". There is, however, nothing inconsistent with the standard models of physics and cosmology in speculating that the actual ensemble is not the "universes" implied by parameterization of these models, but something with greater *a priori* justification. In other words, it may not be the universes themselves that it is appropriate to "pick" from (indeed, I will argue later that there is greater *a priori* justification in picking from something that correlates with the entropy, or information content, of these universes, rather than from the universes themselves).

Smolin believes that we cannot employ a truly random ensemble, because we need something more than just life or consciousness to constrain the ensemble, since otherwise *any* possible life-containing universe would be in our ensemble, and property $\mathcal{W}$ would *have* to apply to our universe, since our universe has life in it, and all (or very nearly all) universes with life have property $\mathcal{W}$.

Thus, $\mathcal{W}$ will never be falsified. But we see now that the assumption made here—that the universes in the ensemble are equally probable—is not justified. In other words, Smolin's **A** assumes that a world-counting statistic will apply.

In the information-theoretic variant of the anthropic principle I am developing, the universes that emerge from our ensemble will have *differing* probabilities, so a random pick from $\mathcal{R}_{\mathcal{L}}$ is not going to favour all life-containing universes equally. Some will be more probable (even much more so) than others. The reason is that we are not "picking" universes. Remember that universes are, in general, synthetic and perspectival entities, and so there is more *a priori* justification in picking from something that is analytically prior, and I will argue that the most justified analytic prior is the notion of an abstract computer program, not the notion of a spacetime universe. I will not go into full details here, saving that for Ch. 6. The main point here is that Smolin's assumptions do not necessarily apply to versions of the AP in which the fundamental entities from which our selection occurs are *not* spatio-temporal entities. This means that Smolin's argument, while valid, cannot be used as the sweeping condemnation of the AP that he seems to imply it is, concluding, as he does, that "not only is the Anthropic Principle not science, its role may be negative," and that

> It must then be considered unacceptable for any theory, claimed to be a fundamental theory of physics, to rely on the Anthropic Principle to make contact with observations. When such claims are made... this can only be considered signs that a theory is in deep trouble, and at great risk of venturing outside the bounds of science. [202]

Of course, the question remains whether Smolin's argument might be generalized, so that it applies just as well if the members of $\mathcal{M}$ are non-spatiotemporal entities. To see that Smolin's argument will *not* survive such a generalization, assume that we start with an ensemble of non-spatiotemporal entities from which the construct of "spacetime universe" is a higher-level emergent property. Assume further that there are multiple entities in this ensemble that give rise to the identical consciousness and personal identity. If we accept the concept of synthetic unity, then from the viewpoint of a particular conscious observer, these various different ontic entities all represent *the same* universe—since he or she, after all, inhabits all of them. Hence, if we view the ensemble as containing universes, we will see some as having higher probability than others, whereas looking at it as an ensemble of ontic entities assigns all the members equal probability.

In such a system, the selection effect is still purely anthropic. Yet, without the equiprobability of *universes,* Smolin's argument does not apply. Since there are some universes in the ensemble that are highly improbable, it could well be that the vast majority of life-containing universes actually have very low probability, in a random pick from $\mathcal{M}_{\mathcal{L}}$, with only a tiny percentage of them having any significant probability. The key point here is that there is no reason to assume that universes (or worlds or observers or observations) are the things that nature provides us to "self-select" from.

It now becomes possible to consider **B´** a compatible extension to **B**, rather than an alternative to it. We do, however, need to make one small change to it:

> To falsify our MWAP-based theory, we need to show that it is feasible in principle that our universe does *not* have property $\mathcal{W}$, where $\mathcal{W}$ is true of the vast majority of universes in $\mathcal{M}_{\mathcal{L}}$ (which is identical here to $\mathcal{R}_{\mathcal{L}}$).

If indeed, the vast majority of universes in $\mathcal{M}_{\mathcal{L}}$ have no significant probability at all (effectively 0%), then it is entirely possible this could be the case. We simply construct a $\mathcal{W}$ that depends on the uneven distribution of worlds. In fact, let's posit (for the sake of argument) that $\mathcal{W}$ is simply the statement "the distribution of worlds is uneven". A random pick from $\mathcal{M}_{\mathcal{L}}$ will almost certainly produce a $\mathcal{W}$-world, if our theory is true, but it is still possible it might not if it turns out that the probability distribution is flat, after all.

We could try to patch up Smolin's assumptions by either*:*

1. dropping the assumption in **B** that the universes in the ensemble are equally probable, or
2. dropping the assumption in **A** that our ensemble consists of spatiotemporal universes.

In approach #1, however, we would have to allow Smolin to raise the question of what it is that permits this unequal distribution, and how we can really call it *mere* self-selection, when we have an uneven distribution with no *a priori* justification. Only by starting with an equally distributed ensemble can we really be sure that our "anthropic" principle is truly a case of self-selection, rather than the result of some unknown physical mechanism that prevents the distribution of universes from being flat. For this reason, I prefer, instead, to take approach #2, generalizing **A** so that it is not restricted to spatio-temporal universes (or anything else that assumes a one-to-one correspondence between phenomenal things and really-existing things):

> **A*** There exists a very large ensemble of ontic entities, $\mathcal{M}$, from which can be inferred (as an emergent property) a very large set, $\mathcal{U}$, of universes (which exist in the same sense that our chairs, tables and our own universe exists), which are completely or almost completely causally disjoint spacetime regions. To the extent that they are causally disjoint, we have no ability to make observations in other universes than our own. The parameters of the standard models of particle physics and cosmology vary over the set $\mathcal{U}$ of universes.
>
> **B*** The distribution of members of $\mathcal{M}$ is random (in some measure) and those members of $\mathcal{U}$ are rare that share with our universe the same standard model of particle physics and cosmology, and the same parameters for those models.

Notice that this does require tweaking Smolin's **B**, as well as **A**, since the randomness mentioned in **B** now applies to a different set of objects than does the statement of rarity (whereas in Smolin's original statement, the randomness and rarity features where ascribed to the same set).

We will allow that the members of $\mathcal{U}$ might actually be the ontic entities of reality, since this is simply the special case of $\mathcal{M} = \mathcal{U}$ (in which case, the universes "emerge" trivially, since they were already there). This special case is, in fact, precisely Smolin's original **A** and **B**, so his version of the AP is still included here, as a degenerate case. Thus, given Smolin's original argument (which I believe is sound), we can conclude that theories based on $\mathcal{M} = \mathcal{U}$ *are* unfalsifiable, in Smolin's sense, exactly as he claimed.

In other words, Smolin's Anthropic Principle is the version of the AP that we get when it is assumed that the ontic entities of reality are the *same* as the phenomenal objects of perception (such as worlds or universes or outcomes). Thus, while Smolin's argument may dismiss many variants of the AP, it by no means dismisses the AP itself as a possible foundation for falsifiable anthropic theories. If we can devise an experiment that could show that worlds are distributed differently from our theory, then we could falsify the theory. There is no inherent reason in principle, that I can think of, to assume that this could not be done—and in Ch. 6-8, I will discuss some (admittedly speculative) ideas on how this might be done, in the context of algorithmic synthetic unity (ASU). And if ASU could be developed thoroughly enough to actually derive the existing quantum postulates from *a priori* principles—as I attempt to do in outline in Ch. 8—it could then be argued that existing quantum experiments *already* count as unsuccessful attempts to falsify an anthropic theory, since quantum theory itself would then become an anthropic theory.

For Smolin, such falsification is not possible, even in principle, but only because he assumes that the anthropic principle mandates a flat distribution of worlds. The astute reader will have long-since realized that Smolin's error is essentially the same as that made by Everett and the frequentist Born-rule provers, as well as most of the Born rule objectors: the assumption that, *a priori*, it is *worlds* or *observers* or *outcomes* that count. And since falsifiability requires there be something other than worlds that count, Smolin concludes that this something-else must come from purely synthetic *a posteriori* (physical) constraints unrelated to the mere existence of life or consciousness, and which hence will be non-anthropic in nature. However, as we have seen, this is not in general a valid assumption. It *would* be valid if *"a priori"* were equivalent to "analytic", and so it is possible (although I am merely speculating here) that Smolin overlooks this hidden assumption because he rejects the synthetic *a priori* as a valid category of knowledge. And given that the rejection of the synthetic *a priori* was a major feature of much twentieth century analytic philosophy and scientific thinking, such a bias is actually quite understandable, from an historical perspective. Nonetheless, it is not a valid assumption to make when the subject of your attack (the anthropic principle) depends so intimately on the synthetic *a priori* category to begin with.

6 Algorithmic Synthetic Unity

6.1 Introduction

In spite of my considerable sympathy for the "What else?" response to the Born rule objection, I feel that there are several good reasons that the objectors are (to at least an extent, justifiably) more sceptical of the Born measure than they are of the Lebesgue measure used in classical statistical mechanics:

1. Quantum mechanics really *is* very much more removed from common experience than classical statistical mechanics. So, if it is *slightly* less clear how classical particles in statistical mechanics impinge on us to cause perception (compared, say, to picking marbles out of a bag), it is *very much* less clear how the particles of quantum mechanics so impinge on us. Thus, the objectors have a right to be more sceptical of quantum probability measures. Extraordinary claims require extraordinary evidence, and many worlds *is* an extraordinary claim.

2. Even if *we* do not consider branch-counting to be obvious, it is still considered reasonable (even self-evident) to many others (including Everett himself), and so amplitude-counting cannot be considered the only alternative on the table. Thus, simply asking "What-else?" without further justification is preaching to the choir.

3. Amplitude-counting does not explain the counter-intuitive nature of probability interference. If amplitudes represent a count of the objective ontic entities that underlie the wavefunction, then why do they interfere with each other? What does a negative or complex amplitude represent? Since this is unclear, we are left feeling that we need a deeper understanding of what amplitudes really are, and what they represent, if we are to accept amplitude-counting as an *a priori* choice for probability calculations. What we need is a reasonable story to tell ourselves about how counting some ontic entity or other will yield interfering counts, for the purposes of probability statistics. Everettians may suggest that we just accept interference on its own terms, since this is what the wavefunction gives us, but this does not give us a way of understanding how an amplitude could both be a *count,* for statistical purposes, *and* interfere. Until we can explain this, it will always seem more intuitive to some people to count some quantity derivative from the amplitudes, that is more intuitively a countable, like worlds or observers.

Born rule objectors have, I believe, spent an unfortunate amount of energy criticizing the failed frequentist proofs, when in fact the whole frequentist branch-counting project is misguided to begin with. A cogent Born rule objection must present a reasonable challenge that cannot be dismissed with "What else?" or a look of patronizing bemusement. I will take the following question to be

the main challenge I am trying to address in this chapter, given that amplitudes are complex-valued and thus interfere with each other, and given that the Born rule amounts to amplitude-counting:

Question. *How is it possible for entities that interfere with each other to be the countable entities for a probability rule, and can programs function as such countables?*

This is not meant to resemble the actual challenge as stated by the Born rule objectors themselves. It is rather what is left of their objections that I consider to be most substantive, after eliminating those claims and arguments which I have already (I believe) shown in previous chapters to be insubstantial. I have decided to focus on the interference feature of amplitudes, as this seems the quality most at odds with a classical count in probability theory, and the best reason for considering that we cannot simply assume amplitude-counting.[54]

In other domains of probability theory, if there are two alternative ways the same event can happen, we count "2" towards our probability calculation (normalize the count, and we have our probability). But in quantum theory, it seems as if, when there are two ways for the same event to happen, then sometimes we count "2", but other times we might even count "0", because one of the events happens with a negative (or, more generally, complex) amplitude, so that the final count is "$1 + -1 = 0$". But what does that mean? In classical probability we do not have such "complex counts" (interference effects) in our probability calculations. So, although it is clear to the detractors why outcome-counting or observer-counting can make sense for a branching multiverse (even if they cannot prove it), it is not clear how amplitude-counting can make any sense at all, partly because it is so obscure what an amplitude would represent in a probability distribution, and why there should be these complex counts and interference effects.

6.2 Observer-generator Indifference

From the generative view of probability put forward in Ch. 4, and the principle of synthetic unity, we can posit a new kind of indifference principle:

Assumption. *Principle of Observer-generator Indifference:*

Each (possible)

(i) generator of conscious observer states, and

(ii) conscious observer state with the same generator

are equiprobable (whether or not the states are identical).

[54] I need to thank Wojciech Zurek, who first pointed out to me reasons along these lines why the "what else?" response was not adequate, as such (although he is not responsible for the particular interpretation I have applied to his comments).

Many, if not all, observation probabilities can be viewed as a special case of observer-generation, although this would often be obviously contrived, as in the marble-picking examples, since there is exactly one way that each possible conscious state could be generated by the mechanism in question. However, for cases involving self-location or self-selection, observer-generator indifference becomes important, since there may be more than one way to generate the *same* conscious state. Thus, while it is common to imagine that it is observers that one counts, it is actually observers *weighted* by the number of ways to generate that observer (the number of "generators") that matters.

The distinction could be construed as mere semantics, since if there are two ways to generate a conscious observer state, one could simply point out that, really what this means is that there are *two* observers (not one), who just happen to indistinguishable for our purposes, and so we are still counting observers, and a simpler principle of observer indifference would be adequate. In the marble-picking case, for example, assume there are three red marbles and seven blue marbles in the bag, yielding $p(red) = 30\%$ and $p(blue) = 70\%$. Now, in a very simplified physical model, we might be able to assume that the conscious state resulting from pulling out and looking at any red marble is identical to the conscious state resulting from pulling out any *other* red marble. It is the same mental state, but there are *three* different ways to generate it. Thus we do not count it as one observer-generator, but three. There are *three* observer-generators, but only *one* observer. If this simplified model is just a coarse-graining of the real world, reflecting our current state of knowledge, then there really are three distinct observers, after all. However, we should still be able to calculate probabilities as if there were only one observer. What possible reason would there be for the epistemic probabilities, given an incomplete model, to be different from what the ontic probabilities would be in a universe where the incomplete model was all there was? This is the reason that the *epistemic* Sleeping Beauty puzzle can be applied to Everettian *ontic* probabilities.

Imagine the simplified physical model really *is* a complete model of the world (so the universe really consists of only one room, a bag, marbles and a very simple observer—assume it is a robot that could be programmed to have only a limited number of possible conscious states). It would follow from synthetic unity that all three possible red picks really *are* one observer, and the seven blue picks another. There really are only two observers, in this case. So when we employ synthetic unity, conflating observer-counting and observer-generator-counting is a real problem. There are two possible observers, one has a 30% chance of being realized, the other a 70% chance. Yet, observer-counting would tell us the probabilities should be equally 50%. Note that in this case, it is not even valid to say there are really three different marbles, since nothing about the laws of physics in this universe would allow the observer to make such a distinction, so it would not be likely for the observer to categorize the three balls as three individual balls; he would just say there were two

balls, one red, one blue, and that laws of physics declare the blue ball as more likely to emerge from the bag.

But note, when we say "each possible way of generating", that a "possible way" can either be a "*merely possible* way" (that could have happened but did not), or an "actual way" (that really did happen). A merely-possible way is one that happens in an alternate possible world, that was not actually realized. But recall from Ch. 4, that it does not really matter if we imagine that such possible ways were *actually* realized off somewhere in this world where they do not interfere with the other possibilities, or if they are merely theoretical alternatives to what actual happens.

For instance, imagine that instead of there being different *possible* ways to generate these conscious states, that we have a machine that actually *does* generate them all, by building rooms and constructing robots to place in them, each with a bag containing a *single* marble. The machine builds ten rooms. In three there is a robot with a bag containing a red marble. In seven, there is a robot with a bag containing a blue marble. There are still only *two* possible conscious states—and hence observers—resulting from the marble pick. But one of these observers is weighted more heavily than the other, because we are counting observer-generators, not observers. Of course, we could just change our semantical perspective and declare that there are clearly ten observers (we could simply define "observer" this way). However, that really only makes sense because we are looking in from outside, and can see that there are clearly ten things. The observers in this world would not see it this way. They would declare that there are two possible observer states after the marble pick, and a *single* observer before. Perhaps they would say that before, the marble was in a superposition of states. However, there is no notion of probability interference in this universe, no entanglement, and no quantum weirdness to alert the observer to the idea that all the possibilities might be real. So, more than likely, even if all the possibilities *were* real, the observer would just assume there was some built-in non-determinism to the universe, rather than opting for the idea of multiple worlds.

Of course, there aren't *really* multiple worlds here. Just multiple rooms. But wait: if this is the *entire* universe, and the rooms are non-interacting, then, from the perspective of this observer, these really *are* alternate universes. The main point here is that the principle of observer-generator indifference does not care whether the different possibilities are "real" or merely "possible". Indeed, why should it make a difference? So long as the different possibilities are mutually exclusive (a criterion for the indifference principle) and therefore do not interact with or affect each other, then we can always take "merely possible" ways of generating something, and "make them real" by building another room or universe isolated from the present one, in which the possibility happens. The probabilities, as we saw in the Sleeping Beauty example, work out the same.

6.2.1 Sleeping Beauty and Indifference

What all this means is that Elga's principle of indifference must be rejected, since it declares indistinguishable mental states to be equiprobable with each other, but only if they are in the same (possible or actual) world. Observer-generator indifference is, rather, equivalent (or at least very similar) to Elga's alternative Absurd-Principle-of-Indifference-that-I-do-not-support. It should be noted, however, that Elga does not argue rationally against this absurd principle, but merely holds up its absurdity as evidence of its untruth. However, as we have seen, in any situation where we are counting indistinguishable conscious observer states (i.e. generated observer states), it does not matter whether they are generated in different locations in the space-time of the same universe, or in different space-time universes altogether. The counts for the purposes of probability are the same.

We can see this in the million-mornings and million-robots versions of the Sleeping Beauty puzzle (§5.1.1): the probabilities should work out exactly the same whether we decide that Sleeping Beauty is to be re-awakened a million times (generated at different times within one universe), or that a million robotic copies will be made (generated at different locations within one universe), or that she will be re-awakened just once in one universe, with a hidden ID tag labelled according to the flip of a fair coin (*i.e.,* with the different labels generated in different *possible* worlds).

The key point is how many possible ways there are to generate the conscious state in question, not whether the generations are real or merely possible. But this way of talking can still lead to problems, especially when, as in the million robots example, there is a mix of actual and real possibilities. In the million-copies version, we might be tempted to say that we have a million and two possibilities, because we have a million copies plus the original in one possible world, and just one uncopied original in the other (this would lead us to the erroneous Thirder solution). Equivalently, in the million-sided-coin version, we have all the observer states in their own separate possible worlds, and again we have a million and two, since we have two possibilities generated by the first head-tails coin flip, and then one of those bifurcates into a million and one possibilities.

However, if we employ observer-generator-counting, this analysis is incorrect. The construction of the million copies, whether actual robotic copies or merely possible ones via the million-sided coin flip, are part of the same generation process as the tails result of the first coin flip. The two coin flips are not mutually exclusive possibilities, and therefore cannot be separated from each other as separate generators. The second coin-flip is a sub-generator of the first, and specifically of one branch of it. Thus, if we come to know that the other (heads) branch is eliminated as a possibility, *then* we can consider the entire generator to be a tails branch only, and all the conscious states are now equiprobable.

By this reasoning, there are only *two* consciousness generators, and their two observers must

count as equiprobable. The second (tails) one, as it turns out, can be further subdivided into a million and one generated observers due to the second coin flip, which are sub-generations. So how do we count sub-generations? And why should they be counted differently? They should be counted differently because they are not true generations, created by independent generators—not any more than Steps 3 and 4 of a cake recipe constitute a true stand-alone recipe. Recall the motivation here is to have a concept of counting in probability theory that works when the ontic entities (the things-to-count) are not necessarily in one-to-one correspondence with appearances (the-things-we-see). The ontic entity here is whatever produces the whole history up to the point where the conscious state in question emerges. The million and one copies were created by two coin flips, not one.

Hence we have part *(ii)* of our observer-generator principle, from which we deduce that *sub-generators must be weighted by the probability of their parent generator.* This in no way conflicts with part *(i)* of the principle. Part *(i)* of the principle counts observer-generatORS, *not* observer-generaTIONS. The reason is simple: a generator is understood as an ontic entity—the actually existing thing that creates/constructs/generates the consciousness (or, in the case of epistemic probability, that which *would be* such, if the model used were a complete model of reality). Hence, if three conscious states are constructed by the *same* generator(s), there is still only one generator, not three. These three states are still, by our principle, counted as equiprobable, since they share the same generator(s). Hence if all three are generated by generators #123 and #345, then they are both equiprobable, since they share the same generator count of 2. However, if one (call it robot A) is constructed by generator #123, and the others (call them B and C) are both constructed by generator #345, then all three will *not* be equiprobable, according to the observer-generator principle. All else being equal, the two #345 observers will each have a probability of 1/4, while the #123 observer will have a probability of 1/2.

Imagine a huge machine with a big "#123" label on the side. It digs and saws and hammers until it has built a robot, with a big letter "A" on the its back. Now imagine another machine, which builds *two* robots, labelled "B" and "C". If the machines (generators), *not* the constructions (generations), are the ontic entities (the things that really exist), then we can only treat B and C as distinct entities if we are permitted to ignore A, otherwise they are really just aspects of the one entity. So, while they count as two observers, B and C, they count for only one observer-generator, call it "BC" (using observer labelling, but this is still the same #345 machine). Now A happens to be a single observer *and* a single observer-generator. "BC", however, generates two observers, but is only one observer-generator. By part *(i)* of the principle, A is equiprobable with BC, and by part

(ii) B is equiprobable with C. All else being equal, this yields

$$\begin{aligned} p(\mathrm{A}) &= 1/2 \\ p(\mathrm{BC}) &= 1/2 \\ p(\mathrm{B}) = p(\mathrm{BC})p(\mathrm{B}|\mathrm{BC}) &= \frac{1}{2} \times \frac{1}{2} = 1/4 \\ p(\mathrm{C}) = p(\mathrm{BC})p(\mathrm{C}|\mathrm{BC}) &= \frac{1}{2} \times \frac{1}{2} = 1/4 \end{aligned} \tag{6.1}$$

where $p(x|y)$ means the probability of x given only generator y.

This is exactly the situation in Sleeping Beauty, and indeed this analysis yields the Halfer solution of David Lewis, *not* Adam Elga's more popular Thirder Solution. The calculus used to compute observer probabilities from observer-generator program counts can be summarized as

$$p(\mathrm{O}_1) = p(\mathrm{O}_1|\mathrm{O})p(\mathrm{O}) \tag{6.2}$$

where $\mathrm{O} \equiv \mathrm{O}_1 \vee \mathrm{O}_2$

Note that this is just an instance of standard conditional probability:

$$\begin{aligned} p(\mathrm{O}_1) &= p(\mathrm{O}_1|\mathrm{O}_1 \vee \mathrm{O}_2)p(\mathrm{O}_1 \vee \mathrm{O}_2) \\ p((\mathrm{O}_1 \vee \mathrm{O}_2) \wedge \mathrm{O}_1) &= p(\mathrm{O}_1|\mathrm{O}_1 \vee \mathrm{O}_2)p(\mathrm{O}_1 \vee \mathrm{O}_2) \\ p(\mathrm{A} \wedge \mathrm{B}) &= p(\mathrm{B}|\mathrm{A})p(\mathrm{A}) \end{aligned} \tag{6.3}$$

More generally, we have

$$p(\mathrm{O}_k) = p(\mathrm{O}_1 \vee \mathrm{O}_2 \vee \cdots)p(\mathrm{O}_k|\mathrm{O}_1 \vee \mathrm{O}_2 \vee \cdots) \tag{6.4}$$

In the million-copies version of Sleeping Beauty, we have a million and two observers, but still only two observer-generators, corresponding to heads and tails. WIth observer-labelling, we call the first generator "A", but for the second we need to concatenate a million and one symbols, so we'll just label it with the abbreviation "BC*etc*". This yields:

$$\begin{aligned} p(\mathrm{A}) &= 1/2 \\ p(\mathrm{BC}\,etc) &= 1/2 \\ p(\mathrm{B}) = p(\mathrm{BC}\,etc)p(\mathrm{B}|\mathrm{BC}\,etc) &= \frac{1}{2} \times \frac{1}{1000001} = \frac{1}{2000002} \\ p(\mathrm{C}) = p(\mathrm{BC}\,etc)p(\mathrm{C}|\mathrm{BC}\,etc) &= \frac{1}{2} \times \frac{1}{1000001} = \frac{1}{2000002} \\ etc. & \end{aligned} \tag{6.5}$$

Note that these probabilities sum to unity and obey additivity.

6.3 Observer-generator program counting and Sleeping Beauty

Equation (6.4) gives us a calculus for probabilities in domains with multiple observers generated by the same entity, and multiple indistinguishable observer states. Most probability applications do not have to deal with both of these issues, and so it is easy to assume that observers are what matter, or just physical entities like "worlds" or "outcomes". The Sleeping Beauty puzzle beautifully illustrates why such assumptions are not universally valid.

Peter Lewis [133] was the first to point out that the Sleeping Beauty puzzle shares the same epistemic structure as Everettian world-branching, and to point out that the analogous Everettian situation demands David Lewis's Halfer solution. The epistemic structure of the Everettian case comes from the actual structure of reality—in fact, it really is an ontic structure, as we are asking about objective probabilities, so it only becomes an epistemic structure if we imagine our observer to have full *knowledge* of reality. In the Sleeping Beauty case, this same structure comes about through Sleeping Beauty's *ignorance* of the full situation she is in, creating observer states that are epistemically identical (but not *completely* indistinguishable). If we assume that what Sleeping Beauty knows about the world is a complete model of her reality, then this structure becomes an ontic structure, and the probabilities become objective. Hence, it would seem fishy for the two scenarios to have different solutions. Peter Lewis's observations, then, provide an effective argument for the Halfer solution *if* one already accepts Everett—or conversely, an argument in favour of the coherence of Everettian probabilities, *if* one already accepts the Halfer solution. For someone who already accepts the Thirder solution, and believes the MWI has serious problems, Lewis's observations are less relevant.

To paraphrase Lewis's main point: even though one structure is ontic and the other epistemic, this should not matter in the least for doing probability calculations. Surely, merely by assuming that Beauty's *epistemically* indistinguishable states are *fully* indistinguishable, we do not change the problem fundamentally. And surely, if we translate these identical observer states into a single quantum system in a superposition of states, this also should not change the probability problem fundamentally—or if it does, then the answer to the Sleeping Beauty puzzle depends on our physical model of reality. The conclusion, then, is that *at least one of* the following must hold for the Sleeping Beauty paradox:

1. The Halfer solution is correct, or
2. The correct solution depends on what theory of physics you adopt.

Certainly, most of those who have weighed in on the Sleeping Beauty puzzle have not assumed that its answer depends on whether they accept a certain interpretation of quantum mechanics,

Everettian or otherwise.

6.4 Observer-generator program counting and Everett

The reason that the Sleeping Beauty puzzle is important in an analysis of the Born rule objection, is that it has the same epistemic/ontic structure as Everettian observation, yet it is an example of this structure in a classical, non-quantum scenario. Thus, if we can settle on an answer to this non-quantum probability puzzle, we should be able to use the same counting method to calculate Everettian probabilities. This will justify our Everettian counting method, independently of quantum considerations *per se*. This is a far superior method of settling on a counting method than simply presuming that it must be world-counting, or observer-counting, just because this seems vaguely intuitive to us.

There are some caveats to this approach, however. One is that there is no strong consensus out there on the solution to the Sleeping Beauty puzzle—this is unlike most probability brain-teasers, like the well-known Monty Hall problem (which has a similar structure, but lacks the self-location/self-selection aspects of Sleeping Beauty). This leaves us with a couple of options:

1. First, we could decide that the Thirder position is true (and it does seem to have majority support). This would mean that the Born rule does *not* follow the expected *a priori* (Thirder) method of counting, and would give additional support to the Born rule objectors.
2. Or, we could decide that the Halfer position is true (and I have tried to make clear in Ch. 5 why I believe that it is). This would mean that the Born rule *possibly* follows the expected *a priori* (Halfer) method of counting, and would give additional support to Everettians against the Born rule objectors.

Note that accepting the Halfer solution does not actually prove that Everettian quantum mechanics demands the Born rule, *a priori*. All we have done is show that there is a reasonable *a priori* counting rule that would *not necessarily* demand world-counting or observer-counting. Hence different world branches can have different probabilities without this in any way showing that the Born rule is a problem for Everett. It still leaves the door open for Everettians to show that observer-generator program counting actually *does* yield the Born rule.

6.5 Observer-algorithm program counting

So far, we have developed a non-computational, non-algorithmic counting rule: the principle of observer-generator indifference. However, given my already-stated commitment to Strong AI, and my long-term research goal of basing quantum probabilities on algorithmic information theory, it is time now to extend the principle into the computational paradigm:

Principle 6.1. *Principle of Observer-Algorithm Indifference:*

Each

(i) algorithm for generating conscious observer states, and

(ii) conscious observer state generated from the same algorithm,

are equiprobable (whether or not the states are identical).

This yields observer-algorithm counting, with a corresponding sub-algorithm rule:

$$p(\mathrm{O}_k) = p(\mathrm{O}_1 \vee \mathrm{O}_2 \vee \cdots)p(\mathrm{O}_k|\mathrm{O}_1 \vee \mathrm{O}_2 \vee \cdots) \tag{6.6}$$

6.6 Observer-algorithm counting and Everett

Given that:

1. Counting amplitudes (the most obvious wavefunction entity to count) works empirically, but
2. It is not clear how amplitude interference could arise in a counting measure, and
3. Counting other higher-level entities, such as worlds and observers, has *not* proved to work empirically, and
4. The wavefunction describes all possibilities, within certain bounds, which possibly even means, effectively, all rationally conceivable possibilities, and
5. Computation likely describes all rationally conceivable structures, and since
6. Probabilities in the wavefunction are relative to observers, in a branching, multi-world structure (assuming Everett), and
7. Observations in such a structure naturally involve self-selection (anthropic) effects, which suggests observer-algorithm counting,

I would therefore submit that we consider the following possibility:

Conjecture 6.2. *The Algorithmic Amplitude Postulate: Wavefunction amplitude-counting is the higher-level result of the more primitive procedure of observer-algorithm counting.*

In this and the previous sections, I have given eleven reasons why this is an elegant *a priori* principle for the interpretation of probability in the wavefunction. Nonetheless, it is important to note that this is far from any kind of "derivation" of observer-algorithm counting, much less of the Born rule itself. Rather, this proposed principle remains highly speculative. It has no empirical support. I am not presenting it here as a falsifiable hypothesis (although neither do I claim it is unfalsifiable). However, I believe I have made a good case for its *a priori* metaphysical elegance, and given that we still do not have an entirely satisfactory account of how probability works in

the Everett interpretation, I think that these considerations make observer-algorithm-counting a promising avenue of exploration.

Note also that just because we have cited maximal rational expressiveness as a positive trait, this does not mean that the ontic entities must necessarily *be* computations. Our conjecture does not quite go that far—although that conjecture has *a priori* merit of its own, being, in a sense, the simplest possible rational conjecture of all. However, we are not here going to demand strict isomorphism between ontic entities and computations, merely that the counting of the former should follow the same statistics as the latter. In other words, we will count entities *as if* they were computations, and see where that leads us.

6.7 Program Counting with Algorithmic Categories

The ontic entities in program counting are abstract programs, not algorithms (in the sense that it is the programs, not the algorithms that we are indifferent toward). The abstract notion of "program" corresponds to the concept of an SK-combinator. The notion of an algorithm is more limited, corresponding only to combinators that are interpreted as solving some problem or other, usually by outputting a result. We can still talk about "algorithm counting", in a sense, since we will be categorizing our programs (in the numerator) as algorithms. For instance, when one calculates the algorithmic probability $p(\varepsilon)$ of a particular conscious mental state x, out of a set of possible mental states Ω, one is essentially employing the classical probability proportion thusly:

$$p(x) = \frac{|x|}{|\Omega|} \tag{6.7}$$

since x's "particularity" as a "particular mental state" is such only phenomenologically (and hence, synthetically), and x is analytically a category or class of program: those that qualify as algorithms for computing a particular mental state. And Ω is also a category of programs, depending on the context of our probability calculation (it might be those programs that qualify as algorithms for any consciousness, or just those that are algorithms for consciousnesses that continue the personal identity of a particular observer's mental state).

Of course, since the Solomonoff probability is calculated as an infinite convergent series, the sample space is countably infinite, and the actual equation is more precisely a limit:

$$p(x) = \lim_{n\to\infty} \frac{|\{i : i \in x \wedge i \in \Omega_n\}|}{|\Omega_n|} \tag{6.8}$$

where Ω_n contains the first n elements of Ω, by the Solomonoff ordering. This limit exists, since the Solomonoff series is convergent; and since it is neither infinite nor zero, there seems no reason to avoid calling this limit a literal probability (there is no reason to believe it to be paradoxical or

ill-behaved as a probability: it is more like the probability that a natural number is *divisible by* 3 than it is like the probability of its *being* 3).

We call the result an "algorithmic" probability, not because algorithms are the fundamental countables of the theory, but rather because we categorize our programs into equivalence classes as algorithms that produce a desired result (in this case, a conscious mental state).

Given Kleene's enumeration theorem, algorithms can *refer to other programs*, even themselves. This point will turn out to be crucial later on. For instance, the input to a program might be interpreted as the encoding for another program, and the output might be interpreted as "Yes" if the inputted program halts, and "No" if it does not. This would be the "Halting algorithm". However, it can be proved that no such computer program exists [217], so the idea of a halting algorithm is nonsensical (although there *is* a notion of "limit-algorithms" that can compute the answers to such uncomputable questions in the limit: see Appendix C). Thus, questions can be conceptually posed that have no computational algorithm to solve them in finite steps.

I will choose to refer to algorithmic probability calculations as either "program counting", or "algorithm counting by program indifference", of which observe-algorithm counting is an example.

So why should an Everettian choose program-counting? Well, the programs we count are observer-algorithms. Like the inputs and outputs of a program, the idea that it produces within its internal state a consciousness is an interpretive view of program, not an inherent part of the program (much the way worlds are an interpretive view of the wavefunction and not, like amplitudes, an inherent part of it). In fact, whatever part of the internal state we consider to be the mind of the observer could be considered the "output" to an algorithm for which the question posed is something like "can you produce a conscious mental state?", or "can you produce a conscious mental state that is Liz, and remembers being Liz yesterday?". There is no input as such for this algorithm; the output might be interpreted as a complete description of the computed mental state, or simply a "yes" or "no" for whether the conscious state was generated.

6.8 A Cosmic Algorithmic Measure

The measure I will adopt is simply $H_S(x)$, the Solomonoff complexity, applied to observer-algorithms. Similar ideas have been put forward by others [149, 212, 214, 87, 191, 190, 139]. Recall from Ch. 4 that (assuming binary bits):

$$p(m) = 2^{-H(m)} \tag{6.9}$$

where m is a conscious mental state, and $H_S(m)$ is the algorithmic information content of m (Solomonoff complexity, more precisely, but I will frequently talk in terms of the Kolmogorov com-

plexity and optimal compressions, most of the time).

Recall that $H_S(m)$ can be understood essentially as the least (or average) number of bits it takes to write an algorithm to output m. The inverse relationship takes us from probability back to information:

$$H_S(m) \quad = \quad \log_2 p(m) \tag{6.10}$$

I call this a "global" measure because it could potentially take into account an entire environment much larger than the organism itself.

Compare this to a local perceptual measure that measures the bits contained in a localized description of the same mental state. This seems to be what Carter has in mind in [42] (based on the work of Dyson [73] and Page [147]). Carter's measure is the total information processed in an act of perception. Of course, this might be construed as potentially taking into account the observer's entire environment, if necessary, in which case it becomes equivalent to a Solomonoff measure (although I see no indication that Carter wishes it to be construed that way). This is a semantic issue: how much of the environment is allowed to be part of the "perceptual act"? Can we include things in the same room? Across the galaxy? At remote times in the past? To be consistent with the principles of ASU, it is important that there be no restrictions on how much of an environment (if any) is permitted. None at all, or an entire vast universe, may be included—whatever is necessary in order to minimize the bit-length of the final program code.

The semantic difficulty with defining a local measure lies in defining "local", an inherently imprecise term. Yet, humans usually think locally, so we need to be able to compare our (very) global measure to more local measures.

Definition 6.3. Define $H_S(m)$ as the Solomonoff complexity of mental state m, and $H_L(m)$ as its "local perceptual complexity", such that
$H_S(m)$ = the average size of SK-combinators that generate conscious state m.
$H_L(m)$ = the average size of SK-combinators that generate conscious state m, and contain primitive projection and/or synthesis operators that correspond to identifiable perceptual or cognitive entities.

The term "identifiable cognitive entity" will not be cleanly defined, but it could represent a neuron, a synaptic weight, a visual process or a higher-level perception—anything at all that can be identified with something mental. These are the conventional terms we use to describe ourselves, but ASU asks us to consider that $H_S(m)$ will be a (potentially) more global measure. So none of the individual operators appearing in the optimal/average representation of m will necessarily have any specifically mental interpretation.

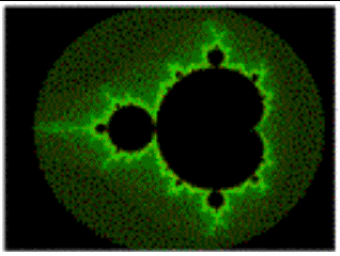
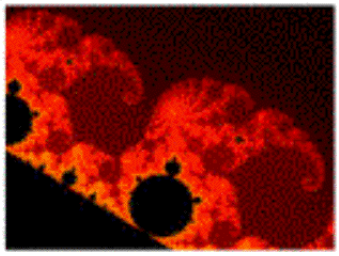
$\{c\} : \lim_{n\to\infty} z_n \neq \infty$ $z_0 = 0$ $z_{k+1} = z_k^2 + c$ $c = x + iy$	 *The Mandelbrot set*	 *A tiny piece of the set*

Figure 6.1: Mandelbrot Set: An example of global simplicity, local complexity

A key feature of $H_S(m)$, in the context of ASU, is the possibility that, if a human consciousness is compressible enough, the information content of the mental state (the perception) may be very much smaller than the local measure $H_L(m)$, yet still pull in the entire physical environment of the observer, including past times, and I will conjecture that this fact has great potential for explaining the order and stability of the universe. Naively, one might assume otherwise, since pulling in more of the environment should lead to a *larger* number of bits, not smaller, and hence lower probability. However, this is not necessarily the case, as I argue in [174]. Tegmark [212] has argued persuasively for a very similar idea. It is entirely possible that, as our measure becomes more and more global, the number of bits will start to reduce at some point, rather than increase. Keep in mind that, in a Big-Bang cosmology, the universe is believed to be unravelling from a simpler state. It may therefore have an informational complexity something like that of the Mandelbrot set (a system of fractals containing incredible detail and infinite variation, yet describable in a very tiny program). If the universe is something like this, in terms of its informational structure, then $H_S(m)$ might produce a very much smaller number of bits than $H_L(m)$, and correspondingly higher probabilities.

Such higher "globalized" probabilities are required if our measure is going to elegantly explain the stability of the universe. The purely local measure of a perceptual event, call it $H_L(event)$, will not be so different from the number of bits, call it $H_L\,(random)$, in a program that simulates the same perceptual event in a near-empty world, with only random inputs from the environment. This would not rule out unlawful or dream-like universes.

I will argue that if this informational measure is to explain the stability of the universe, we should expect to have:

$$1 \ll H_S\,(world) = H_S(event) \ll H_L(event) \approx H_L(random) \ll H_L\,(maverick) \ll H_S\,(maverick) \tag{6.11}$$

where "event" refers to a locally defined perceptual event, such as the perceiving of a measurement result; "random" refers to a perceptual event in a random environment with no particular structure,

and "maverick" refers to a perceptual event in a contrived or "maverick" (rather than merely random) universe. The different measures listed above are given from left to right in order from lowest-bit (highest probability) to highest-bit (lowest probability).

Maverick worlds have low probability because they cannot be described compactly at a global level. In other words, we must resort to a local description, explicitly describing particular brain or perceptual states.

It takes many more bits to describe the maverick perception than the random world. Therefore, it will always be more probable that you are just in a random world, and the maverick behaviour is a fluke. Such an observer could never justifiably believe themselves to be in a maverick world, persisting over time, as they would most likely immediately fall into a more compressible world. This is the same basic idea that underlies Hanson's "world-mangling" [102], and if we use his metaphor, we would say that the more compressible world "mangles" the less compressible one. Note that although these maverick worlds are represented in the wavefunction, they are not inhabitable or perceivable worlds.

By allowing total freedom, in terms of the level of description, in computing $H_S()$, we allow the split between observer and environment to be made at any point that facilitates a maximum compression rate—*i.e.*, we squeeze the perception into the most compact description possible. There is, after all, no clear physical criterion for making an absolute split between where an observer ends and its environment begins. A neurological description is surely unjustified, for this explains the perception in terms of specific physical entities (neurons, dendrites, neurotransmitters) that we hypothesize are producing the perception in question. We could rather determine to always stick to a purely perceptual description, in terms of mental states and their properties, but then this would prevent us from ever explaining anything, and would lead to the kind of complete subjectivism we have been trying to avoid. But if we are to draw the dividing line at neurology, what is to stop us from drawing it further out into the environment, and including a description of the whole body? Or the whole room the person is in, or the whole universe the person inhabits, including all its past states? If we are to stick to a realist, objectivist account of nature, then whatever entities we choose to count, the resulting measure must not depend on drawing an arbitrary observer-environment division line.

The ontic entities, therefore, that we will count to generate probabilities will be the bits in an abstract computer program capable of *generating* a mental state, rather than the bits processed by a perception. We are looking here for the lowest-bit description of the mental state, while allowing this measure to be as local or global as it needs to be to reach a minimum value, whether that means describing just the brain state of the observer, or the entire universe of the observer.

This explains *why* we do not live in a random, unlawful or maverick universe. Such universes will take many more bits to describe than our highly compressible universe, which we are assuming takes fewer bits to describe than even our individual brains. Because $H_S\,(world) \ll H_L\,(random) \ll H_S\,(maverick)$, I will suggest that contrived or maverick worlds should not even be called "worlds" at all. Suppose we imagine, purely as a thought experiment, a person who *already* lives in a maverick world. Granted, this is an unlikely individual to be sure, but surely we can still contemplate this individual, and surely their world is perfectly real and solid *to them.*

But this reasoning does not actually hold up. An observer in a maverick world will, with near certainty, immediately drop right back into a more stable, orderly (compressible) universe, as there are far more programs that generate their consciousness that describe compressible universes than uncompressible maverick ones. Indeed, it is more likely that such a consciousness is hallucinating in a lawful world than that they actually inhabit an unlawful one.

Definition 6.4. The "world" of conscious state m consists of all the phenomenal entities generated by a stable (large) proportion of the programs that generate m. Such stable entities are "physical" or "material", while analogous unstable entities are "possible" or "hypothetical".

6.9 Making Falsifiable Predictions

We saw earlier that Smolin's proof that anthropic theories are unfalsifiable applies only for an ensemble of equiprobable worlds. Using algorithmic information as a probability measure would assign different probabilities to different worlds, without the need for any physical mechanism to constrain the distribution of worlds, since information content is a pure *a priori* mathematical property.

In fact, use of this measure could possibly result in falsifiable predictions, if estimates of information content of the early universe can be made. The local information content $H_L(U)$ of the very early universe U should be more or less the same thing as the global information content $H_S(m)$ of a human mental state m, which should be far less than the local measure $H_L(B)$ of a human brain B.

Conjecture 6.5. *Condition for Cosmic Algorithmic Stability:*

$$\begin{aligned} H_L(U) \approx H_S(m) &\ll H_L(B) \\ p(U) = p(m) &\gg p(B) \end{aligned} \tag{6.12}$$

ASU would be falsified if this condition were shown to be violated by the combined efforts of cosmologists and psychologists.

The condition sounds completely counter-intuitive–why would the program for the whole universe (including my brain) be shorter than that for my brain alone? But, given ASU, it must be the case. If it were not, and my brain on its own had the shorter program, then my brain on its own would be the more probable state to find myself in, and there would be no stable observable environment. Yet, we know from experience that the most probable state to find ourselves in is one of a large, complex and law-like environment, so if ASU is correct, it must take more bits to describe a brain alone than an entire universe.

Note that the uncomputability of H_S is not a large barrier to this kind of falsification, since we do not actually need to compute a Solomonoff series to perform the test, since we can ignore $H_S(m)$ and simply seek to demonstrate the following:

Condition 6.6. *ASU Falsifiability Condition*:

$$H_L(U) \gtrapprox H_L(B) \tag{6.13}$$

both of which are local measures.[55] Cosmologists will simply have to tell us more or less the structure of the initial state of the universe (including its size). Psychologists will have to tell us more or less how to write a program to simulate a conscious human brain (including its size). If the former size is greater than, or even approximately equal to, the latter size, then ASU cannot be true.

6.10 Quantum Theory, ASU, and the Problem of Idealism

The ASU explanation for cosmic stability rules out some of the more fanciful speculations based on the MWI and maverick worlds [146, 155, 174, 138] that have worried some and enticed others. It also provides an answer to concerns that the MWI carries with it all the familiar problems of idealism—once everything thinkable is real, why should there be any order or stability in the world? What of ethics and human decision-making in a world where every possible future happens? [157] John Bell said of Everett's interpretation, "if such a theory were taken seriously, it would hardly be possible to take anything else seriously" [18, p 139][15, pp 123-7].

Such worries are similar to those of Dr. Johnson in his refutation of Berkeley's idealism, wherein he simply kicked a stone and said "I refute him thus" [24]. The fact that the stone reliably kicks back meant to Johnson that there must be more than ideas to the world around him. Indeed, it might not be an exaggeration to say that every defence of materialism or physicalism, against idealism, is ultimately an appeal to "kick-back". The basic idea of algorithmic stability answers this objection tidily, whether in defence of an all-out analytic idealism or of an Everettian multiverse.

[55] However, even if we were concerned directly with the Solomonoff series for $H_S(m)$, its uncomputability would in no way rule out the possibility of a reasonable estimate.

To many of its detractors, the MWI is really just idealism in an empirical guise. However, a measure such as H_S allows us to understand Johnson's "kick-back" in terms of information content and probability, so that maverick worlds simply cannot be inhabited or observed, even if we attempt to factor them into the wavefunction. Thus, there is still a distinction between a possible world (a world that can be consistently imagined) and actual physical worlds (which have the informational stability to support observers), a distinction that many materialists have long presumed was unavailable to the idealist.

Some may argue that Everettianism is not a form of idealism, however, since not *all* possible worlds are necessarily included in the wavefunction, for instance those that violate unitary evolution. However, an argument can be made that any possible *perception* does indeed result from Everettianism. This follows from the fact that any perception that would result from a non-unitary transformation can be duplicated by (a possibly more complex) unitary transformation. And, since worlds are defined synthetically, *ipso facto* all possible worlds are included in the Everettian multiverse, since all possible phenomenal entities have some nonzero probability.

Of course, not all possible worlds are equally probable; however, since they all *do* exist, isn't this a version of idealism? Perhaps not. If not all worlds are equally probable, we have to look at the source of the unequal probabilities. If this source is something purely idealistic and nonmaterial (such as possible perceptions or possible mathematical constructs) then we have a kind of idealism (although what kind depends on the nature of the source). However, if the source of the unequal probabilities is non-ideal, physical or material in nature, then our ensemble is not idealism, in spite of its containing all possible worlds. However, even in this case, one can see why those who do not like the consequences of idealism might also have similar issues with Everettianism. Everett himself, however, seemed to believe (unjustifiably, in my opinion) that all worlds *were* equally probable (at least in the limit), so it would seem that Everett's own view was effectively idealistic, with respect to worlds (or branches). Or perhaps we could say he was a materialist for single cases, but an idealist "in the limit" (for what it is worth, his biographer considers him to have been a materialist [40]).

In any case, it is clear that Everettianism has many of the same uncomfortable consequences as idealism, and when one chooses, as I have, to count programs as ontic entities, rather than something material, it becomes much harder to avoid idealism. This may not bother some, but many others wonder how to account for everything from ethics to a law-like environment to free-will, in a reality that allows such a plethora of possibilities to see the light of existence. Several thought experiments that Everettians have used to justify bizarre consequences—like manipulating the probabilities to win the lottery, or being immortal—have exacerbated these concerns (and, in fact, according to his biographer, Everett himself actually believed that his interpretation guaranteed him immortality

[40]).

In the coming sections, I will examine several of these thought experiments, and argue that most, if not all, of the bizarre consequences do not hold up under ASU. As we will soon see, the culprit behind almost all of the bizarre conclusions is, once again, the unjustified assumption that all worlds must be equiprobable.

6.10.1 Thought Experiment: Winning the Lottery

Assume that any (consistent) experience you can dream up exists in your wavefunction. As we have already argued, there is reason to believe this is the case. There is, for instance, a world where all the gas particles in the room you are now in suddenly rush into a corner of the room, suffocating you. This violates the classical laws of thermodynamics, but it happens in such a small percentage of the wavefunction, that the chances of your observing it are mind-bogglingly tiny. There is also a universe where random molecules in the room spontaneously organize into a million dollars in gold, or a hot fudge sundae, or a flying pig. There is another universe where they congeal into a virtual reality (VR) machine that makes you *think* you see a flying pig. And there is yet another one where the molecules in your brain organize themselves so as to make you hallucinate a flying pig. And another where you hallucinate that you have observed a violation of the conservation of energy in a laboratory experiment. But all these highly contrived possibilities pale in comparison to the vast majority of the wavefunction, where you continue to breath the air molecules in the room, as you are right now, and nothing unusual happens at all (of course, this means the majority of the *wavefunction,* not the majority of *worlds*, which are actually the weird maverick ones).

So the question arises: can we not do something to change the probabilities so that the more desirable alternatives become more probable? I will call such feats "miracle-working", since we are trying to violate our usual notions of what is physically possible.

Definition 6.7. *Miracle:* any gross violation of the classical laws of physics.

I will distinguish between several types of miracles.

Definition 6.8. *Micro-miracles*: miracles that are on too small or restricted a scale to have any productive impact on our everyday lives, compared to what was understood to be possible classically.

Definition 6.9. *Macro-miracles*: miracles that have the potential to impact our everyday lives in a productive way. Of these, we will distinguish two further subtypes:

Minor miracles: macro-miracles that do not appear in themselves to violate the classical laws of physics, but which radically shift the probabilities of events in a way not allowed for classically.

Major-miracles: macro-miracles that appear to violate the classical laws of physics.

To give an example, a magic spell that brings love or allows me to win the lottery would be a minor miracle, since nothing that happens would appear in and of itself to violate physics. A levitation spell or invisibility cloak, on the other hand, would be a major miracle (so long as they were not achieved through clever technological or mechanical means).

The quantum lottery experiment is one of the thought experiments used to prove the possibility of minor miracles, given the Everett interpretation. Let's say you are playing a lottery with a 1 in 10 million chance of winning the jackpot. The universe where you win clearly exists in the universal wavefunction, even shortly before the draw. Lottery numbers are usually chosen by a machine that mixes up a lot of small balls with numbers written on them—a process that is clearly chaotic, so that a very small quantum uncertainty will very quickly make a difference at the macroscopic level, resulting in decoherence. Hence, all possible lottery numbers are about equally probable right up to some time fairly close to the draw, and after the draw, you are the big winner in one ten millionth of your wavefunction. To generate a minor miracle, you want to increase the probability of winning from $1/10{,}000{,}000$ to something reasonably close to 100%.

There is potentially a way to do this, at least in theory, but first allow me to insert the following disclaimer:[56]

> The following technique for minor miracle-working is strongly recommended AGAINST. It is a philosophical thought experiment only, intended to stimulate intellectual discussion. It is highly unlikely to work in practice, most probably ending in death or serious injury, even if it "works" in some idealized theory (which is, in any case, highly questionable), and I take no responsibility for anyone unbalanced enough to actually attempt it.

Having said that, here is the technique: simply kill yourself if you do not win the lottery! The best way to do this would be to have a machine automatically monitor the lottery results, perhaps from a website. You then go to bed before the draw, and have the machine quietly perform the execution while you sleep. Since you are never conscious or aware in the universes where you lose, they are automatically no longer part of your physical wavefunction. You can only wake up in a universe where you have won.

One annoying (or pleasing?) consequence of this method for minor miracle-working is that it is only verifiable to the one who performs it. From everyone else's point of view, you have not changed

[56] The origin of this thought experiment is unclear. Its first publication seems to be on USENET discussions in the 1980s. I have had personal communication with the person I believe to have (possibly) originated it, but he or she does not wish publicity, due to the risk of being associated with something that a misguided individual might actually try, with fatal consequences. However, it is worth noting, that this person, like myself, sees it as a purely academic gedanken experiment, believing it to be too rife with complications to be actually practical (even if one placed absolute trust in all of the premises).

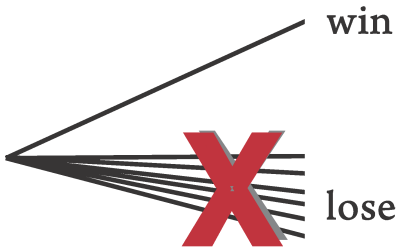


Figure 6.2: Minor Miracle: Winning The Lottery

the odds of winning the lottery at all by killing yourself. If they witness repeated attempts at this experiment (presumably performed by different people!), the result they see will be a dead corpse 99.99999% of the time.

There are several really good reasons not to go out and try this, even if you believe in the MWI, and even if it so happens you do not care about the perspective of others (perhaps you have no family or loved ones). First of all, do you really have enough faith in the MWI to bet your life on it? (I certainly do not!) Secondly, even if you could be absolutely certain of the theory's truth, you had still better be very sure to do your probability calculations properly, and that you set the procedure up properly. If you mess up, the result will not be millions of dollars, but a failed suicide attempt that may leave you crippled for life. For instance, if a gun is used, what is the probability of its jamming and not firing? Or inflicting serious injury, rather than death? In the case of a human accomplice, what is the chance that they will chicken out, or faint? I believe that even leaps from very tall buildings occasionally end in survival.

I would strongly suspect, in fact, that a one in ten million lottery draw would be too difficult to ever pull off in practice, no matter how carefully set up. Even if it *is* doable, one might well have to spend the equivalent of the prize money just setting it up. If one looks at the rate of suicide attempt failures, it is simply *far* higher than one in ten million (and I would wager it is so for just about any suicide method you can think of). Even if one is absolutely sure that the winning number is being chosen in a way that is truly random on the quantum level, it is difficult to imagine a suicide method that one could be confident would work every single time out of ten million cases.

Even one in a thousand would probably not be doable, since you will have less money to spend on the setup if you are to see a profit. Indeed, it is quite likely that *no* size of lottery will be doable in practice, no matter how small, since the money and effort you can afford to devote to the project diminishes with the size of the prize, and the risk of failure that you might be willing to tolerate will be less at lower payouts, as well. This is simple human nature: some high risk-taking people might be willing to perform a dare-devil stunt with a 1% death rate for a million dollar prize, but

no one would take such risk for, say, a tiny ten dollar prize. Imagine, for example, a $10 door prize at a small cocktail party of ten people. The lottery miracle might (superficially) seem doable at this level, since you only have to come up with a suicide method with a failure rate sufficiently lower than 10%. How much lower depends on your tolerance for risk, but let's say you settle on (a still *very* risky) one order of magnitude ($1/10$) as "sufficiently lower", so that you will require a suicide method with a failure rate no worse than 1%. While it is true that suicide methods that are 99% successful are hardly miraculous, typical suicide success rates are still much lower than this, and you have almost no money to spend on setting this up, and very little time to spend working on the project, to achieve a reasonable return for your efforts (you might have an hour to plan and set up, just to earn minimum wage). Indeed, for such a small prize, you are unlikely to take this much risk, so you will probably need to do much, much better than this. Even ignoring the uncertainties and subtleties of relying on the MWI, the odds, the risk involved, and the payoff just do not add up to an attractive gamble, even for a relatively risk-taking personality.

So we will conclude that the lottery miracle is most likely *not* doable in practice, even if one is very generous about the assumptions required for it to work. Nonetheless, it still seems doable in some sense, at least "in theory" as a gedanken experiment—especially if we are willing to ignore the numerous practical problems, such as the requirement to actually turn a profit. So let's accept the result of this thought experiment, at least for argument's sake, and we will find that it leads us into even more bizarre territory.

6.10.2 Thought Experiment: Invoking a Magic Elf

Now that we have devised a method for minor miracle-working (even though the results are unverifiable to others, and most probably impractical), we will try to apply the same technique to create major miracles. Let's say I wish to invoke a tiny magical green elf to appear before me in a puff of smoke. This possibility necessarily exists somewhere in the wavefunction, just as did the lottery win. The only difference is that the probability is even lower—surely a difference in degree, not in kind. Thus, I ought to be able to apply the quantum suicide trick, killing myself in any universe where a tiny green magical elf does *not* appear before me in a puff of smoke. This time, I will use a human accomplice to detect the presence of the magic elf while I sleep, since the detection procedure will probably involve physical examination, interviewing, testing of the elf's magical competence, and so on. If I am not being too fussy, the test of minimum magical competence could simply be the elf's ability to appear out of nowhere in a puff of smoke.

Unfortunately, there are numerous difficulties with this scheme, which we will now examine.

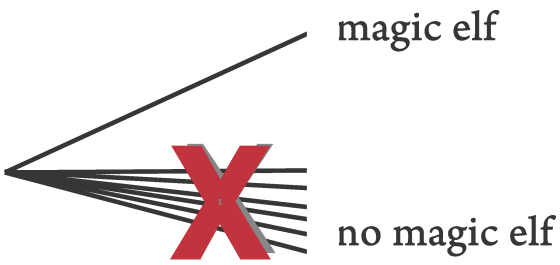


Figure 6.3: Major Miracle: Invoking A Magic Elf

Reliability of Detection While my accomplice might be perfectly able to recognize a tiny magical green elf if such a being were to appear, this does not at all imply that he can detect that one has *not* appeared. Just because he believes there is a magical elf in the room, does not mean that there is one, and that it will continue to be there, acting lawfully like a magic green elf in the future when I wake up (whatever the laws governing magical green elves are!). For instance, my friend might be hallucinating an elf, in which case, when I wake up I will see nothing unusual at all (other than an apparently hallucinating friend). Or, perhaps a second friend heard about the vile experiment, and is trying to save my life by faking the appearance of a magic elf, complete with dry ice, magic tricks and special effects. Both of these possibilities, and probably many more, are almost certainly *far* more probable than a real, probabilistically stable, tiny magical green elf appearing in a puff of smoke.

This problem was not so insurmountable with the lottery miracle, since in that case it was arguably possible (even if difficult) to detect a winning lottery number accurately enough that the probability of incorrect detection was sufficiently lower than the probability of having a winning number.

Reliability of Suicide Method Essentially the same problem we had with the detection method reappears in the suicide method itself, as it did for the quantum lottery. This made the lottery miracle infeasible in practice, but it renders the magic elf miracle impossible, even in theory. Forget here about spending millions of dollars setting up a sure-fire suicide method. Even something as unlikely as a bullet's magically vaporizing just before it hits you is probably *far* more likely than the magical elf you are trying to create. You would be much better off trying to breed a kennel of dogs to *evolve* into magic elves!

Failure Conditions We can lump the detection and suicide reliability problems together into a single process. Assume that $p(ideal)$ is the probability of the ideal actually happening, and $p(\neg ideal)$ is the probability of the ideal not happening. In the absence of the ideal, the probability of

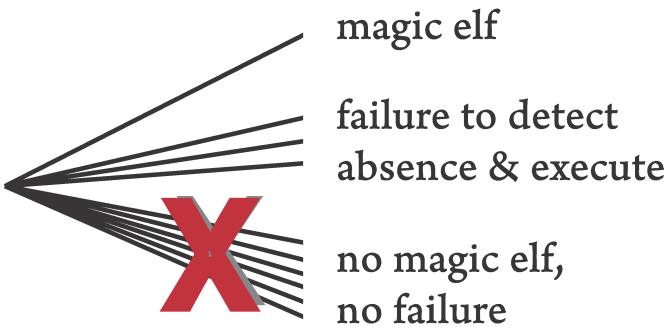


Figure 6.4: Intractability of Major Miracles (Invocation Miracle)

successfully detecting the absence and committing suicide is $p(suicide)$ and the probability of failing to do so is $p(\neg suicide)$.

$$\begin{aligned} p(suicide) + p(\neg suicide) &= 100\% \\ p(ideal) + p(\neg ideal) &= 100\% \end{aligned} \tag{6.14}$$

From this, we can state the condition for failure of miracle generation:

$$\begin{aligned} p(\neg ideal) &\gg p(suicide) \\ p(\neg suicide) &\gg p(ideal) \end{aligned} \tag{6.15}$$

While the lottery experiment may be doable in theory under these standards, even if not in practice, the magic green elf experiment is certainly unachievable. Even if we drop the requirement for the elf to be consistently magical, his appearance is sufficiently improbable to be considered effectively impossible.

6.10.3 Thought Experiment: Universal Immortality

So minor miracles are perhaps possible, at least in some idealized theoretical sense, even if they are impractical and publicly unverifiable. But more dramatic miracles, which actually seem to accomplish the impossible, are probably just that: impossible. But what about the miracle some would consider the most desirable of all: immortality? One might think this is surely a major miracle, and thus as impossible as the magic elf. But this is not so obvious.

Immortality is a very special kind of miracle, with surprising properties. What happens if I try the lottery trick here? That would mean that I kill myself in all worlds where I do not survive. But wait... this is already done for me, by definition! No consideration need be given here to methods of suicide and their probabilities of failure, since mere survival is all we care about. As long as I do

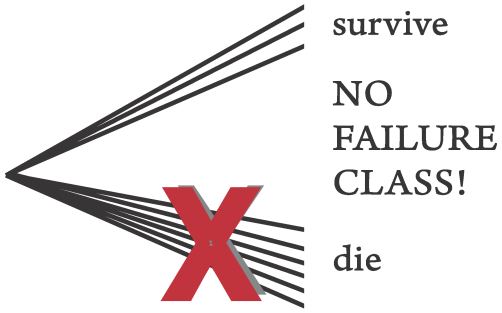


Figure 6.5: Universal Immortality: A Tractable Major Miracle?

survive in even a tiny, tiny percentage of the wavefunction, then immortality is automatically mine. This means that we are, all of us, *already* immortal, at least from our own individual perspectives. No further argument is required. Let's celebrate!

Or so goes the reasoning.

This is the basic structure of the synthetic-unitary immortality proofs, all based on Strong AI, put forth by people like Moravec [146], Tipler [215] and Perry [155]. A probabilistic analysis of such proofs yields a kind of "least miracle required for salvation" rule: if a miracle is required to save us from death, then the most probable such universe is the one with the least miraculous method of salvation. Unlike arguments for other kinds of major miracles, synthetic-unitary immortality proofs based on Strong AI are not so easy to dismiss. However, they are not so clear-cut—or as attractive—as some of their advocates seem to believe, as we will soon see.

Accidental Death If I get run down by a passing bus, there will clearly be some universes in which I survive–at least from the point of view of some time shortly before the accident occurs. At the point when major miracles would be required to save me (such as the bus levitating in the air, just missing me), it is possible that my death becomes 100% fated to happen. But even so, my doomed state will not last for very long. Given the generally chaotic nature of the world, it seems a safe bet that this doomed state will not be reached until shortly before the accident (just as in the lottery draw). Thus, so long as I can live with the fact that I might, in a small percentage of universes, exist for brief periods of time in a doomed state, then minor miracles seem to offer some measure of immunity from accidental death.

To some extent, this is a matter of attitude. I can, if I choose, view these brief fated periods (which are bound to exist in some worlds) the same way I view short periods of memory loss within a single universe. Many of us have experienced short-term memory loss, such as from an accident or drinking binge. We do not generally consider the loss of a few hours to be a form of death. Likewise,

why should we fret over the fact that we will be briefly fated to die, in some small percentage of future worlds? Of course, this is no excuse for recklessness, since getting hit by a bus could still leave you in pain and suffering for a prolonged period of time.

As with the lottery, this is not an effect that is publicly verifiable. But unlike in the lottery example, this effect is not even, in general, subjectively verifiable for particular cases, since these occur accidentally, rather than being intentionally set up as in the lottery example. So even you will not know when such minor miracles have happened to you.

Aging Of course, for those who seek immortality, what really matters is whether we can conquer aging, the cause of the inevitable death that seems to await us all. But for that, we may have to invoke major miracles. Imagine I am 98 years old, on my deathbed, and the best modern medicine can keep me alive no longer. It seems, on the face of it, unlikely that continued survival should be *logically* impossible. Surely, in some tiny percentage of the wavefunction, events will conspire to keep me alive at least a little bit longer, even if major miracles are required. Unfortunately, if this is what is in store for me, I had best be very afraid, as it will lead to a state of eternally increasing decrepitude, in which I spend the rest of eternity undergoing a series of ever-more-miraculous "least miracles required for salvation", just barely keeping me alive [174]. This sounds less like my ideal of immortality, and more like some Struldbruggian Hell. Okay, this is not what I had in mind by "immortality". Help!

David Lewis saw this Struldbruggian result as a major reason, not to discount Everett's claims, but to fear them. "Everett's idea is elegant, but heaven forfend it should be true!" he said [131]. (Interestingly, Lewis also supported the Born rule objection against the many worlds interpretation, on the basis of *a priori* branch-counting, in spite of the fact that he also supported the Halfer solution to Sleeping Beauty, which I have argued is antithetical to branch-counting.)[57]

Tipler and Perry seemingly avoid the Struldbruggian problem by suggesting that the most likely way for me to survive is not the "least miracle required to keep me decrepit", but rather a miracle that essentially causes me to quantum tunnel to a spatially and temporally disconnected high-tech resurrection, either in the far future or a completely different universe! This occurs because, in some universe, high-tech future beings (or perhaps aliens) will build a robotic replica sufficiently like me that the synthetic unity of consciousness forces us to conclude that it *is* me.

[57] Leslie [126] has constructed an argument against Everett that is based on very similar ideas. Instead of arguing that the MWI implies eternal decrepitude, he argues that the number of versions (branches) of a person very close to death will be very much greater than the number that are not, meaning that by the anthropic principle, we should expect to find ourselves close to death. The fact that we are not is thus a puzzle in need of explanation. This version of the argument, however, clearly has no wings once we have rejected branch-counting as an *a priori*, since it explicitly invokes branch-counting. On the other hand, as we will see shortly, the Struldbruggian argument is really just the same argument in disguise, since it will ultimately fail due to the invalidity of branch-counting, as well.

How do the high-tech beings achieve this feat, in spite of the fact that I have been long dead (or in the case of the alien universe, never lived at all)? Perry suggests that they simply build the best replica they can, and fill in the rest with a lucky guess. Tipler suggests some (slightly) more plausible scenarios, but admits they might not work, and so ends up resorting to our being accidentally recreated by sheer enumeration, given the MWI—which of course returns us right back to the Struldbruggian problem.

We are clearly flirting here with predictions that violate classical thermodynamics on a massive scale, so we should be *very* sceptical of these results from the outset. The Struldbruggian option would seem far more likely, but would also clearly violate classical thermodynamics.

But let us now apply the tools of algorithmic probability. A contrived universe in which my 98-year old self is miraculously saved has a very long program, perhaps not even having the same history as the real universe (it might even be a simulation in a computer inside an even larger universe). So all we have to do, in order to refute both the eternal decrepitude *and* the brute force resurrection theories, is to show that there exists some other kind of universe that also continues my consciousness, but whose shortest program size is on the same order of magnitude as the information content of just my brain, B. Such a universe would be much more probable (have a shorter program) than the highly contrived miraculous universes mentioned above, and hence would be much more probable (but not nearly as probable as the "real" universe, based on my mental state m).

We can construct such a universe by coding a program that simulates my brain without providing a spacetime environment for it. We replace the spacetime environment with a virtual reality (VR) of negligible program size, which feeds the brain essentially random input. This sounds a lot like dreaming. In the Crick-Mitchison model of dreaming, for instance, the brain stem floods the brain with random noise, and the brain free-associates a dreamworld in response, in order to erase spurious information [58]. But a dreamworld has little stability and is not very law-like in its behavior. In addition, there would be little reason for this replica brain to be competent at storing long-term memories—or even storing short-term memories with any competence or reliability—as such robustness would likely not be necessary to simply continue my stream of consciousness. Such a brain might also have many other unfortunate defects (remember, this is the least required miracle, and there is no alien or high-tech intelligence in this random-VR world to make sure that my brain is constructed in a reasonable fashion). Hence, calling such a world a "dream-world" is probably already extremely generous.

Let us call the random-VR universe R, for "random". The guesswork universe and/or the eternal decrepitude universe we will call C, for "contrived". R is much less probable than m, essentially

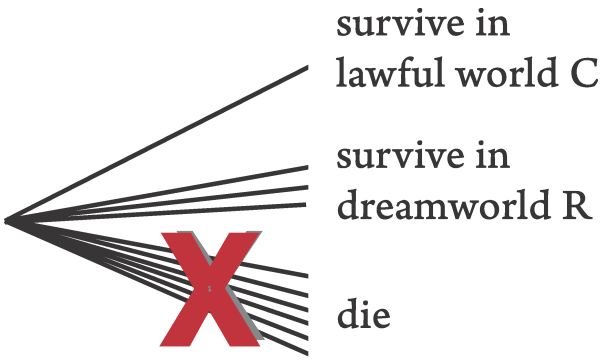


Figure 6.6: The Intractability of Universal Immortality

equal in probability to B, and much more probable than C:

$$\begin{aligned} H(m) \ll H(B) &\approx H(R) \ll H(C) \\ p(m) \gg p(B) &\approx p(R) \gg p(C) \end{aligned} \tag{6.16}$$

So there is a sense in which the MWI does guarantee that a future resurrection world exists, tailored just for me, out there somewhere in the universal wavefunction—such a world is a logical possibility, and is therefore accounted for in the wavefunction, contributing to it some tiny amplitude. But this does not actually make me immortal, since I am much *more* likely to quantum jump into a dream-like state than into such a futuristic paradise. Such a paradise, unlike the real universe, lacks stability. Thus we are, I believe, justified in refusing to even call it "physical", a "world" or a "universe".

Even if I fantasize about being that incredibly lucky, logically possible chap who *does* quantum jump into a futuristic robotic replica, this still places that version of me in a universe with vastly lower probability than the random VR, practically guaranteeing an *immediate* quantum jump back into the random VR again. This is why it makes no sense to even say that this fellow was *ever* "in" such a world. And in the random VR, not only does the environment lack lawful behavior, but it is doubtful that one's brain would even function properly or lay down memories in any kind of reliable fashion; remember that this is the *least* miracle required for continued consciousness. There would be little distinction, in fact, between various random so-called "worlds". Remember, according to algorithmic synthetic unity, it is *because* $H(m) \ll H(B)$ that there is an arrow of time in a time-symmetric mechanics, which ensures that it will be very unlikely to have recoherence (reverse collapse or wavefunction merger). But here, there is no such principle, and no reason at all to presume there will not be just as many mergers as collapses! The different "worlds" would lose all distinction... they would all tend to blend into each other, time "flowing" in both (or any) direction, and it would be difficult to define any coherent long-term history for any such "worlds"

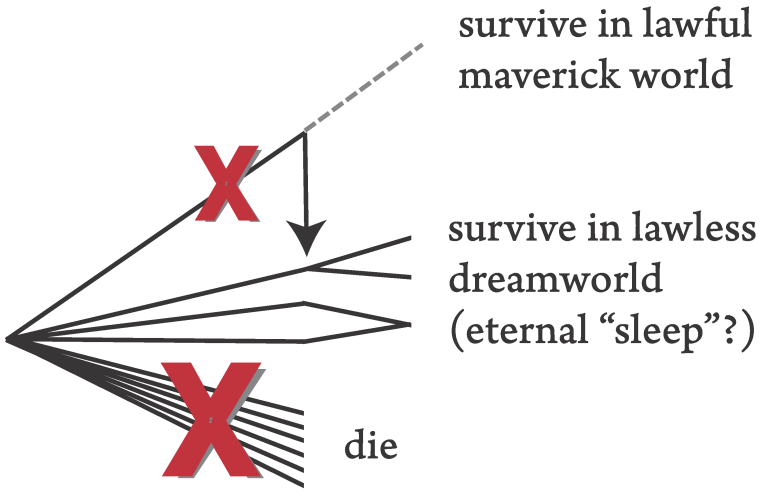


Figure 6.7: The Instability of Universal Immortality (via World "Mangling")

as being distinct from the others. And all this is still probably being generous in our assessment of these "worlds", given that there are probably lots of other issues that such a "world" would have that we haven't even mentioned or thought of. I think it is not too presumptuous to suggest that living in such a random, incoherent world of transience would be less like living and more like a state of eternal sleep—if indeed the idea of such a state is even coherent enough to lend this much credit to.

In short: the future me who lies on his deathbed, with no immortality medicine to save him, is as good as dead.

6.10.4 Thought Experiment: Personal Immortality

So quantum major miracles have turned out to be illusory as a means to immortality, even of the Struldbruggian kind. This is not so surprising, since any quantum immortality argument requires violations of classical thermodynamics on a massive scale (although I have nowhere assumed the validity of classical thermodynamics in refuting them). Are we destined to die then? Perhaps not. With advances in anti-aging and related biotechnologies—including dietary supplements, genetic manipulation, therapeutic cloning, embryonic stem cell research and nanotechnology—perhaps we can avoid death through the actual sweat of our brow. Of course, even if such medical breakthroughs come along, we would still eventually succumb to accidental death, so it seems that this would provide us only with indefinitely prolonged life, not *literal* immortality (in the sense of eternal life).

Except that we have already seen that there is an argument for quantum immunity from accidental death, which might imply immortality in combination with anti-aging technology. Unfortunately, there is still a flaw here if you are looking for strict immortality [138, 174, 214], since *any* argument

for strict quantum immortality will founder on the simple fact that any such immortal observer will, at some point, have their Born measure reduced to the point that their world becomes a maverick world—in other words, their continued existence would be a major miracle.[58] Imagine, for instance, that (due to advanced technology) you are free from aging and disease. You reach a point in your life where being killed in an imminent accident becomes, by the quantum equations, a 99% certainty—perhaps a bus is barreling down on you at a cross-walk. No problem for your continued survival, you say, since 1% is still a significant quantum probability—unlikely, but clearly not in the major miracle category. The problem is that, if you are looking for strict immortality, there is no limit to the number of such 1% minor miracles you will need to survive, over the long haul. While surviving one accident with only 1% chance may not be a major miracle, perhaps surviving a large number of such potential accidents would be—and it would seem that at *some* point, the continued consciousness of our so-called "immortal" observer would surely slip into the major miracle category, and hence be "mangled" out of existence.

Remember, it is ultimately the entire situation that matters (the Born probability of the entire sequence of accidents or near-accidents, the whole history). Once the measure of this history goes below a certain threshold, and its probability becomes competitive with that of maverick worlds, there is no tenable way to maintain that you have "survived", since the so-called "world" you have survived in is "mangled" (in Hanson's sense), having no order or "kick-back". Any "experiences" such a survivor might be imagined to have will have a higher Solomonoff probability of being part of some *other* more orderly world, where they are simply hallucinations or madness.

While it still might be possible that an argument for quantum immortality could be made to work, given some more detailed analysis, short of that, I think the idea can be, for the time being, dismissed. As an aside, it is significant to note that, while Everett appears to have had a *personal* belief in quantum immortality [40], there is no reason to believe that he thought he had a viable argument for it, as he never published on the subject. On the other hand, insofar as his published theory tries to maintain both wavefunction realism *and* world-counting—something I am claiming is untenable—I would argue that Everett's published position (particularly his stage 2 Born rule proof) sets him up for belief in quantum immortality. The antidote is algorithmic probability theory, which gives us a tenable philosophical justification for setting priors according to an indifference principle that is radically different from counting observers.

[58] Credit goes to Jacques Mallah for convincing me that no form of the quantum immortality argument is currently viable.

7 Toy Examples

7.1 Counting Algorithms: a toy example

We will now examine some toy examples of observer-algorithm counting, and how it can be used to calculate probabilities. I will list several very simple programs (written in BASIC-F pseudo-code, which is hopefully self-explanatory, but consult Appendix D for further details). We will assume this small set of programs makes up the entire ontology of our system. At this stage, the goal is to set up the simplest of toy systems, simply to illustrate how certain phenomena—such as interference effects—can come about in such a system. The programs will not realistically reflect the structure of any actual minds or worlds.

Next, we imagine that some very simple pattern within these programs constitutes an emergent conscious state (this is the output of our algorithm). We will not use a realistic criterion for consciousness (that will be something for future work). Next, using the algorithmic definition of information content, H_S, and the principle of synthetic unity, we will look at how to calculate actual probabilities.

Note that these programs are not intended to be (even toy) programmatic versions of quantum states. The idea here is specifically *not* to impose the structure of quantum dynamics on our system. We are attempting to show, rather, how a computational system, with observer-algorithms as the countable entities, can display some of the basic features of quantum theory. The assumption of the Born rule objectors (and Everett) is that something like observer-counting is the correct general *a priori* method for calculating probabilities for observers emergent from a formal system. We wish to show that this is not the case, and that there are such systems in which world counting does not work, and in particular, there are such systems in which probability interference is expected. Thus, to assume quantum mechanics would be to defeat our purpose in this chapter (although I will attempt to sketch out a path back to real-world quantum mechanics in the next chapter).

The purpose of this chapter is to explore the features of observer-algorithm counting, and hopefully convince the reader that it reproduces many of the basic features of quantum theory in a way that is at least consistent with amplitude-counting in quantum mechanics and inconsistent with

branch-counting.

Let's start with a few simple programs:

Toy Example #1

```
Program P1{1}       // H = 1 bit
Let a=a+1
Let a=a-1
LOOK(a) // nothing
// no observer

Program P2{10}      // H = 2 bits
Let a=a+1
Let a=a-2
LOOK(a) // 1

Program P3{1100}    // H = 4 bits
Let a=a+1
Let a=a+1
LOOK(a) // 2

Program P4{1101}    // H = 4 bits
Let a=a-1
Let a=a-1
LOOK(a) // 2

Program P5{111}     // H = 3 bits
Let a=a+4
Let a=a-2
LOOK(a) // 2
```

where:

- the "//" introduces a comment, which has no functional significance,
- all variables, when not explicitly assigned a value, take on a default value of 0,
- arithmetic is on the integers,
- the BASIC-F code shown is the uncompressed program, with the optimally compressed binary version in curly braces {} after the program name. H values shown in comments are the bit counts of the compressed programs. No actual compression algorithm has been run on these programs; the binary numbers are simply made up.
- a conscious observer "Liz" is present whenever there is a nonzero value for the variable a:

$$\text{Liz's consciousness} \Longleftrightarrow a \neq 0 \tag{7.1}$$

- the `LOOK(a)` command causes Liz to observe the magnitude of variable `a`. So that Liz has some limit to the resolution at which we can examine variables in her world, we will imagine that Liz's brain is not equipped to recognized the difference between positive and negative numbers, so that 2 and -2 are both seen as "2". The "//" introduces a comment, which has no functional significance; so in "`LOOK(a) // 2`", the "2" is just there to note the phenomenological result of the `LOOK` command from Liz's perspective (*i.e.,* it shows you what she actually sees).

7.2 Synthetic histories

Definition 7.1. A "merger" or "superposition" $P_{i_1}+\cdots+P_{i_n}$ of (the programs in) set $\{P_{i_1},\ldots,P_{i_n}\}$ is a single program that computes all the "component" programs in the set and no others. We will also use $P_{i,j,k,\ldots}$ as a shorthand for $P_i+P_j+P_k$, and we can use P_{i-j} for a merger $P_i+\cdots+P_j$ of consecutively indexed programs P_i-P_j.

Definition 7.2. A "(consistent) synthetic history" or "world" is a set of programs $\{P_{t_i}\}$ with partial ordering $P_{t_i}\rightarrow P_{t_j}$ such that each program P_{t_i} is a merger of all programs that generate a particular conscious mental state m_{t_i}, and where m_{t_j} is a continuer for m_{t_i}. Each t_i is a "moment" in "synthetic time" for the "person(s)" defined by this history.

Definition 7.3. Two programs are "consistent" if there is at least one conscious mental state that they both generate.

Now let's draw a probability tree for the programs in Example #1. This yields probabilities at the leaves of the tree, by looking at which programs we have the power to describe, as we acquire more and more bits of information about the system. For instance, if we receive a single bit, 0 or 1, the best this does is tell us either that we are dealing with P_1, or one of the others; and so on (see Fig. 7.1). We will define an additional infinity of maverick programs P_{1a}, *etc.*, which are not listed with our other programs because they produce "Liz" states, but are extremely long, so they can for all practical purposes be excluded from our program count.

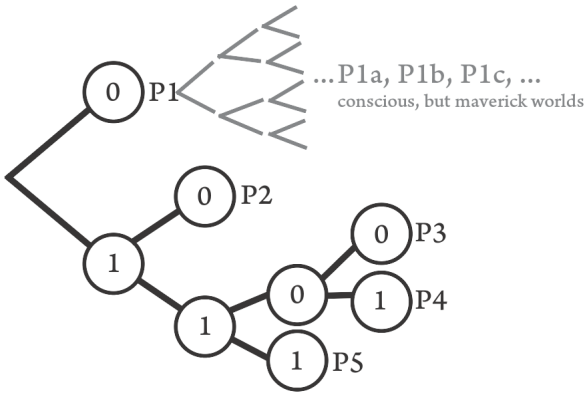


Figure 7.1: Probability Tree for Toy Example #1

The probability tree in Fig. 7.1 yields the following entropies and probabilities (with check-marks indicating which possibilities to include in the count):

$H(P_{1a,b,\ldots}) =$ many bits:	$p(P_{1a}) \approx p(P_{1b}) \approx \cdots \approx 0$	✗
$H(P_1) = 1$ bit:	$p(P_1) = \frac{1}{2^1} = \frac{1}{2}$	✗
$H(P_{2-5}) = 1$ bit:	$p(P_{2-5}) = \frac{1}{2^1} = \frac{1}{2}$	✓
$H(P_2) = 2$ bits:	$p(P_2) = \frac{1}{2^2} = \frac{1}{4}$	✓
$H(P_{3-5}) = 2$ bits:	$p(P_{3-5}) = \frac{1}{2^2} = \frac{1}{4}$	✓
$H(P_{3-4}) = 3$ bits:	$p(P_{3-4}) = \frac{1}{2^3} = \frac{1}{8}$	✓
$H(P_3) = 4$ bits:	$p(P_3) = \frac{1}{2^4} = \frac{1}{16}$	✓
$H(P_4) = 4$ bits:	$p(P_4) = \frac{1}{2^4} = \frac{1}{16}$	✓
$H(P_5) = 3$ bits:	$p(P_5) = \frac{1}{2^3} = \frac{1}{8}$	✓

Essentially, what we mean by "synthetic history" is a path (history) that we trace through the running of all possible programs, which must trace out a single stream of consciousness. If there is more than one program that produces the conscious state in question, then we merge them into one program, for the purposes of answering questions about *that* particular conscious state. This program merger we can also call a "superposition" of programs. If the two programs produce different histories, then we have a superposition of entire histories. We could also call synthetic histories "consistent histories". So long as two programs produce the *same* consciousness, there is no inconsistency in merging them, but if they produce *different* consciousnesses, then we cannot merge them, for to do so (thereby calling them a single history) would be to claim that two different consciousnesses can be the same person. In the domain of algorithmic synthetic unity, this practically is the definition of inconsistency. (I am not claiming equivalency here to the consistent histories interpretation of quantum mechanics [104, 98], since this is not quantum mechanics, but there are obvious parallels.)

Note that a single history may define more than one person, since the same program may generate more than one consciousness. If so, we say these two people "inhabit" the same universe. One program continues another so long as *one* of its mental states continues *one* of the mental states of the other.

Note also that while a history is defined as partially ordered (so there may be more than one program for the same synthetic time), in the real world, histories appear to be more or less totally ordered. However, partially ordered histories are possible in principle. What if two programs both produce the same consciousness *now*, but yesterday, that same person was a superposition of two people; there is no inconsistency there. It doesn't seem to happen in the real world, but it is not impossible to imagine such a thing: "Today, I am looking at a spin-up result, but yesterday I can remember looking at both spin up and spin down at the same time, on this same machine. As I recall, at the time, neither one of us was aware of the other, but now I can clearly remember both." This would be an example, in quantum mechanics, of wavefunction merger or recoherence,

the time-reverse of wavefunction collapse (recall from Ch. 2 that it is unlikely to happen in the real world, but is not mandated against by any rules of quantum theory).

Keep in mind, also, that the shorter programs will always predominate in the probability count, and so if we merge two programs into one, even if they are consistent with each other, but one is much longer than the other, then for all practical purposes, we might as well leave the longer one out.

As an example, the merger $P_3 + P_4$ could only appear in a history if P_3 and P_4 are consistent with each other, which they are, as they both produce the mental state of perceiving a "2". In fact, to include only P_3 and *not* P_4 would be inconsistent. If they produce the same conscious state, it would never be possible for someone to claim that she is in P_3 and not P_4: she is necessarily in superposition. If a continuer mental state distinguished between P_3 and P_4, then we might include one without the other for that synthetic time.

If merged programs are informationally independent of each other, then merger amounts to running them in parallel (multi-tasking). Of course, there could be overlap between the two programs—mutual information—in which the programs could be re-written, with communication between the programs. It doesn't necessarily take as many bits, in other words, to specify a merger as to specify its parts independently (and this can be see in our probability tree).

A crucial point, which the reader should be able to see by now, is that programs, while they can serve as ontic entities, have no built-in "wall" separating them. Any two programs can also be considered as one program, and indeed, the entire set of all programs can be combined into a single, super-program (essentially, the combinator that enumerates and runs all combinators). We need to build "walls" for synthetic, rather than analytic reasons.

7.3 Calculating Probabilities

We begin our synthetic history construction by listing *possible* histories (including inconsistent ones):

$P_1+P_2+P_3+P_4+P_5$	$P_1+P_2+P_3$	P_4+P_5	P_1
$P_1+P_2+P_3+P_4$	$P_1+P_2+P_4$	P_3+P_5	P_2
$P_1+P_2+P_3+P_5$	$P_1+P_3+P_4$	P_2+P_5	P_3
$P_1+P_2+P_4+P_5$	$P_2+P_3+P_4$	P_1+P_5	P_4
$P_1+P_3+P_4+P_5$	$P_1+P_2+P_5$	P_3+P_4	P_5
$P_2+P_3+P_4+P_5$	$P_1+P_3+P_5$	P_2+P_4	

$$\begin{array}{ll} P_2+P_3+P_5 & P_1+P_4 \\ P_1+P_4+P_5 & P_2+P_3 \\ P_2+P_4+P_5 & P_1+P_3 \\ P_3+P_4+P_5 & P_1+P_2 \end{array}$$

From an Everettian/Thirderist world-counting point of view, all of these possible histories would be "equally likely", since they are all "equally real", and they should all have a probability (before renormalization) of

$$p = \frac{1}{N},\ N = \text{number of histories} \tag{7.2}$$

In our example,

$$p = \frac{1}{31} \approx 3.2\% \tag{7.3}$$

But, of course, different programs can have different information contents, and hence different probabilities, so if the notion of program or recursive function is ontologically priori to that of a "world", we should not expect world-counting to be the way to calculate probabilities. (Most of these histories, in fact, have zero probability, since only a few are consistent with Liz's current consciousness.)

There are only two conscious states generated by all five programs: 1 and 2. So there are really only two possible stable *synthetic* histories:

$$\begin{array}{l} P_3 + P_4 + P_5 \\ P_2 \end{array} \tag{7.4}$$

There are still, of course, a large number of unstable (but not inconsistent) synthetic histories, such as P_{1a}, which we could include if we wished to be more precise (and the resulting probabilities would be affected by a tiny amount). However, we are choosing to weed these out from the beginning, so the probabilities of our two remaining histories are:

$$\begin{array}{rl} p(P_{3-5}) & = \frac{1}{2^{H(P_{3-5})}} = \frac{1}{4} \\ p(P_2) & = \frac{1}{2^{H(P_2)}} = \frac{1}{4} \end{array} \tag{7.5}$$

Renormalizing to account for self-selection, we have

$$\begin{array}{rcl} p(P_{3-5}) & = & \frac{1}{2} \\ p(P_2) & = & \frac{1}{2} \end{array} \tag{7.6}$$

This looks like a world-counting result, since we have equiprobable worlds, but that is mere coincidence, as the two branches here just happen to have the same information content. It is easy

to modify the example slightly, to get unequal probabilities, by simply changing one of $P_3 - P_5$ so that it does not generate a conscious mental state.

Toy Example #2

```
...
Program P5{111}    // H = 3 bits
LOOK(a) // nothing
Let a=a+4
Let a=a-2
```

Notice that all we did here was move the `LOOK` command to the beginning of the program, so that a is 0 and no consciousness results. Of course, a is still 2 at the end of the program, just as it was before. But since Liz's "brain" (the variable a) never looks, it does not see the 2, and since seeing a nonzero is a requirement for this "brain" to be conscious, no conscious state is generated by this program at any time.

Adjusting our raw probabilities:

$$\begin{aligned} p(P_5) &= \frac{1}{8} \\ p(P_{3,4}) &= \frac{1}{8} \\ p(P_2) &= \frac{1}{4} \end{aligned} \tag{7.7}$$

and renormalizing with a normalization constant of $1/8 + 1/8 + 1/4 = 1/2$,

$$\begin{aligned} p(P_5) &= \frac{1}{4} \\ p(P_{3,4}) &= \frac{1}{4} \\ p(P_2) &= \frac{1}{2} \end{aligned} \tag{7.8}$$

we get three worlds, one of which is twice as probable as either of the other two, contradicting the world-counting statistic, which would demand $p = 1/3$ for all three.

7.4 Collapse

So far, a history is just a single program, as we have only one conscious state (2), so there is no way to consider whether some states are continuers of other states. So there is only one moment in a synthetic history, and thus far a history consists of a single program. In example #1, P_2 was thought of as a single program and P_{3-5} as a merger or superposition of three programs. However, this is clearly an arbitrary artifact of how we decomposed the programs, since P_{3-5} is really just a

single program. We could just as well have decomposed P_2 into "three" components and left P_{3-5} as a single BASIC-F program.

Neither view is any better in the current situation, since we only have a single moment of time, and no possibility of collapse[59]. However, as as we add additional synthetic moments, it is possible that we might have multiple continuers for a single continued state.

Definition 7.4. A "collapse" or "split" occurs when there is a member of a synthetic history P_k that can be decomposed into multiple component programs, $P_{k_1}, P_{k_2}, \ldots, P_{k_n}$ where each component generates a unique continuer of the same conscious state m.

Note that nothing so far has mandated that the continuer be generated *by the same program* as the continued state. This is to be expected, since how we break things down into separate programs is arbitrary. Under some decompositions, the continued state might be generated by the same program as the continuer, and in others it might not (however, we will see as we go along, that it would be unlikely to have a non-maverick continuer that does not also generate its continued state).

Let's modify example #1, then, to allow a further time step, and therefore a collapse of $P_{3,4,5}$. We already have enough lines of code for two time step, we just need to provide an opportunity for observation in between the two assignments of a.

I will gloss over the issue of continuers here. Our model of Liz is not complex enough to allow any sensible test for continuers, so to simplify the analysis, I will assume that (in our examples thus far) any conscious state that appears in a later time step is a continuer for any that appeared in an earlier one in the same program. Assume that we have performed the relevant tests that have already shown the states to be in the synthetically correct order (later, though, we will shuffle up this ordering, and of course, will not be able to assume that correct ordering has been maintained).

Toy Example #3

```
Program P1{1}      // H = 1 bit
Let a=a+1
LOOK(a) // 1
Let a=a-1
LOOK(a) // nothing

Program P2{00}     // H = 2 bits
Let a=a+1
LOOK(a) // 1
Let a=a-2
LOOK(a) // 1
```

[59] Sometimes, even for a proponent of many-worlds, the word "collapse" with its Copenhagen connotations, can be superior to the word "split". Since the collapse/split event is merely perspectival, one could argue that "collapse" is better because it describes better what happens from the observer's perspective.

```
Program P3{0100}   // H = 4 bits
Let a=a+1
LOOK(a) // 1
Let a=a+1
LOOK(a) // 2

Program P4{0101}   // H = 4 bits
Let a=a-1
LOOK(a) // 1
Let a=a-1
LOOK(a) // 2

Program P5{011}    // H = 3 bits
Let a=a+4
LOOK(a) // 4
Let a=a-2
LOOK(a) // 2
```

Now let's again look at the consistent synthetic histories:

At moment 1:

P_5

$P_1 + P_2 + P_3 + P_4$

At moment 2:

$P_1 + P_2 + P_3 + P_4 \rightarrow death$

$P_1 + P_2 + P_3 + P_4 \rightarrow P_2$

$P_1 + P_2 + P_3 + P_4 \rightarrow P_3 + P_4 + P_5$

$P_5 \rightarrow P_3 + P_4 + P_5$

In the first universe, Liz essentially dies, since after the second program step, there is no longer any pattern in P_1 that represents her personal identity. In the next two universes, the superposition collapses, and Liz observes a "1" and a "2", respectively. The fourth universe is quite odd. Here Liz goes from a P_5 universe—and P_5 is independent of P_3 and P_4—to suddenly being back in a superposition with $P_3 + P_4$ again. This is not a splitting of the world, but a re-merging of worlds (although recall that this is theoretically possible even in real world quantum mechanics). If we take a close look at our toy model, we can readily see why it happens here. P_5 coincidentally takes Liz's mental state back to being identical to that which it has in $P_3 + P_4$. This coincidence happens here without too much contrivance, because we have a toy model with a single variable representing a person's entire mental state. If the real world does work this way, the actual pattern that constitutes Liz's identity will be incredibly complex, resulting from the massively parallel interactions of billions of complex neurons. The number of bits that it would take to describe such a state would be much greater than 1, so coincidentally matching it in another universe with a significantly different history

would seem much less likely.

So let's modify our example again, to eliminate this unlikely, coincidental artifact, by making P_5 do something different to Liz from the other programs, keeping everything else the same.

Toy Example #4

```
Program P1{1}      // H = 1 bit
Let a=a+1
LOOK(a) // 1
Let a=a-1
LOOK(a) // nothing

Program P2{00}     // H = 2 bits
Let a=a+1
LOOK(a) // 1
Let a=a-2
LOOK(a) // 1

Program P3{0100}   // H = 4 bits
Let a=a+1
LOOK(a) // 1
Let a=a+1
LOOK(a) // 2

Program P4{0101}   // H = 4 bits
Let a=a-1
LOOK(a) // 1
Let a=a-1
LOOK(a) // 2

Program P5{011}    // H = 3 bits
Let a=a+4
LOOK(a) // 4
Let a=a-1
LOOK(a) // 3
```

And now our possible consistent histories are:

At moment 1:

P_5

$P_1 + P_2 + P_3 + P_4$

At moment 2:

$P_1 + P_2 + P_3 + P_4 \rightarrow death$

$P_1 + P_2 + P_3 + P_4 \rightarrow P_2$

$P_1 + P_2 + P_3 + P_4 \rightarrow P_3 + P_4$

$P_5 \rightarrow P_5$

When will the superposition collapse? As soon as the observer's conscious state encodes information about other variables in its component program that distinguishes between the superposed programs (*i.e.,* when the observer is correlated with an appropriate "environmental variable").

Note that, in general, use of the `LOOK()` command does not "disturb the system", in the sense that the observable, a, is not affected by our looking at it. I mention this only to make clear that we are not building into our example anything like a disturbance principle, where one cannot observe something without modifying it in some way. Nonetheless, we have ended up with a kind of disturbance principle, anyway, on the level of synthetic histories. The collapse occurs because Liz looks at (or more precisely, becomes correlated with) a. Liz can reasonably tell herself, then, that if she *hadn't* looked, her universe would still be in the superposition. But its not that looking disturbs the system, with the laws of Liz's universe somehow preventing her from looking without disturbing. Rather, the apparent "disturbance" happens because Liz gathers information about which universe she is in, and it is this very fact—of her conscious state's encoding information from her environment that distinguishes between the universes in superposition—that causes (thanks to synthetic unity)—a collapse into one of those universes.

Now let's calculate the probabilities involved in these histories. Note that the history involving Liz's death is not really a separate consistent history, since her death results in no more consciousness, and that universe is thus no longer under consideration. This doesn't mean that we can't talk about what happens in it, and ask questions about it (the tree still falls in the forest even if there is no one left in the universe to see it fall).

The calculations for both pure and empirical self-selective probabilities are shown below, for each history.

History	t	H	*Empirical probability*	*Pure probability*
P_5	1	3	$.125/1 = 12.5$	$.125/1.375 = 9.1\%$
$P_{1,2,(3,4)}$	1	.2	$.875/1 = 87.5\%$	$.875/1.375 = 63.6\%$
$P_{1,2,(3,4)} \rightarrow P_2$	2	2	$.25/.375 \approx 66.7\%$	$.25/1.375 \approx 18.2\%$
$P_{1,2,(3,4)} \rightarrow P_{3,4}$	2	3	$.125/.375 \approx 33.3\%$	$.125/1.375 \approx 9.2\%$
$P_5 \rightarrow P_5$	2	3	$.125/.125 = 100\%$	$.125/1.375 \approx 9.2\%$

The empirical probability, recall, tells us the chances of ending up with a given continuer state given a prior continued state. The is the kind of probability that could be tested by simply repeating observations many times over, and recording the statistics. This is Liz asking herself, "Given that I already know I am Liz, and I am making this observation, what are the probabilities of the possible outcomes?" The pure anthropic probability is very different: it is the probability of *this* history

being selected out of all other possible histories. This is Liz asking herself, "What are the chances of being *me* here now, instead of being someone or somewhere else, possibly in some other universe?" Note that the pure probabilities do not sum to 100% because the first and last histories are actually the same program.

7.5 Interference

My main goal in this section is to show how we can get interference effects in algorithmic synthetic histories, to remove the mystery over how it is that probability counts can possibly interfere with each other. The main problem is to understand how two ontic entities (for us, that means programs) can each, if considered on their own, count positively towards the probability measure of a subsequent conscious state, yet when we consider that *both* are possibilities, they cancel out, so that there is no possibility of the mental state happening at all. Of course, we don't need to choose an example with *total* destructive interference, but this will prove the point most dramatically, and presumably if we can create a scenario that displays complete destructive interference, existence of similar systems with partial interference should follow.

Question. *How could the existence of* two *programs that compute the* same *conscious state, end up meaning the result is* less *probable, even impossible? No matter what* P_4 *does, doesn't* P_3 still *generate Liz's consciousness?*

Actually, it is possible that it will not, if P_3 and P_4 can *refer to each other*, effectively allowing them to mutually stomp on each others' memory space—in other words interfere with each other! Since Liz's consciousness consists of any nonzero in memory location a, and P_3 would generate Liz's consciousness (were P_4 ignored), then if P_4 were to undo the work that P_3 did to generate Liz, and P_3 were to do the same to P_4, then each will have interfered with the other's ability to generate Liz, and there is no generative capacity (amplitude?) for Liz remaining. Yet, each, if it were not for the other, *would* generate Liz!

Destructive interference.

Note that such interference is not possible if the programs are mutually independent of each other; if $H_S(P_3 : P_4) = 0$. To further develop this, we need to create a new command for our pseudo-code language that allows program reference:

$$\mathrm{P}k.var[t] \Leftrightarrow \text{ variable } var \text{ in program } P_k \text{ after program step } t.$$

Note that "program step" t is not necessarily the same thing as "synthetic time" t. The former refers to computation steps in the low-level program, which may or may not correspond to the latter, which refers to synthetic time steps for the observer.

Since none of our reasonable candidates for an analytic basis language give us the ability to directly change what happens in one program from another program, we cannot really have one program directly stomping on another program's memory (and this "stomping" is perhaps somewhat metaphorical, like Hanson's "mangling" [102]). Pk.a can only be used to *reference* P_k's memory, not actually modify it. However, by mutually referring to each other, the programs can affect each other in an indirect manner.

To see how this can happen, we will change our example once again:

Toy Example #5

```
Program P1{1}          // H = 1 bit
Let a=a+1
LOOK(a) // 1
Let a=a-1
LOOK(a) // nothing

Program P2{00}         // H = 2 bits
Let a=a+1
LOOK(a) // 1
Let a=a-2
LOOK(a) // 1

Program P3{0100}       // H = 4 bits
Let a=a+1
LOOK(a) // 1
Let a=a-P4.a[1]
LOOK(a) // 0

Program P4{0101}       // H = 4 bits
Let a=a-1
LOOK(a) // 1
Let a=a-P3.a[1]
LOOK(a) // 0

Program P5{011}        // H = 3 bits
Let a=a+4
LOOK(a) // 4
Let a=a-1
LOOK(a) // 3
```

Our consistent synthetic histories are now:

At moment 1 *(unchanged)*:

P_5

$P_1 + P_2 + P_3 + P_4$

At moment 2 *(interference between P_3 and P_4)*:

$P_1 + P_2 + P_3 + P_4 \rightarrow death$

$P_1 + P_2 + P_3 + P_4 \rightarrow P_2$

~~$P_1 + P_2 + P_3 + P_4 \rightarrow P_3 + P_4$~~ *death*

$P_5 \rightarrow P_5$

Renormalizing, and re-calculating probabilities yields the results below.

History	t	H	*Empirical probability*	*Pure probability*
P_5	1	3	$.125/1 = 12.5\%$	$.125/1.25 = 10\%$
$P_{1,2,(3,4)}$	1	.2	$.875/1 = 87.5\%$	$.875/1.25 = 70.0\%$
$P_{1,2,(3,4)} \rightarrow P_2$	2	2	$.25/.25 \approx 100\%$	$.25/1.25 \approx 20.0\%$
~~$P_{1,2,(3,4)} \Rightarrow P_{3,4}$~~	~~2~~	~~3~~	~~33.3%~~	~~9.2%~~
$P_5 \rightarrow P_5$	2	3	$.125/.125 \approx 100\%$	$.125/1.25 = 10\%$

We see that P_3 and P_4 together result in Liz's death, because they cause a to be zeroed. Or perhaps we should say she has been "cancelled". It might seem strange to talk about someone dying because they were cancelled out by wavefunction interference, since such things do not seem to happen like that in the real world, but keep in mind that this is an exceedingly simple example, where the entire personal identity of a human being is being modelled by a single on-off bit of information, so it is really not unrealistic here to speak this way.

In any case, the main point is not so much that Liz "dies", but that there are no legitimate possibilities represented by P_3 and P_4 in the synthetic history ensemble—but only because P_3 and P_4 *both* co-exist in the ensemble of programs. If we were to imagine eliminating, say P_4, then (presuming that 4 is an index to an enumeration of programs) P_5 would now be P_4 and we would have:

Toy Example #6

```
...
Program P3{0100}     // H = 4 bits
Let a=a+1
LOOK(a) // 1
Let a=a-P4.a[1]
LOOK(a) // 3

Program P4{011}      // H = 3 bits
Let a=a+4
LOOK(a) // 4
Let a=a-1
LOOK(a) // 3
...
```

The synthetic histories are:

At moment 1 *(unchanged)*:

P_4

$P_1 + P_2 + P_3$

At moment 2 *(interference between P_3 and P_4)*:

$P_1 + P_2 + P_3 \rightarrow death$

$P_1 + P_2 + P_3 \rightarrow P_2$

$P_1 + P_2 + P_3 \rightarrow P_3 + P_4$

$P_4 \rightarrow P_3 + P_4$

And now P_3 is back to "counting" again. Only in combination with the old P_4 do we get the mutual interference that causes both to go to zero.

7.5.1 The Legitimacy of Program Reference

A possible objection to this explanation for interference is that the program-to-program reference we have used is not "real reference". Isn't it kind of cheating, after all, to simply, by fiat, give programs (our ontic entities) the right to mysteriously refer to other ontic entities (just by defining a notation that *by definition* refers)? What, after all, does this notation "`P4.a[1]`" actually *do* computationally?

Strictly speaking, this is true enough. Programs do not inherently "refer" to other programs. To see that they do not, remember our discussion of algorithms and programs in Ch. 4. Programs, as ontic entities, must be independent of any *particular* programming language they may be written in. Rogers' isomorphism and the analytic Church-Turing thesis give us a fairly high confidence that programs really *are* objective entities. Yet, most programming languages have elements that not all (and perhaps very few) other programming languages have. The ability to refer to other programs would be one such feature, since it does not exist in most languages, and certainly not in the SK-calculus. Basic arithmetic is shared by almost all (but certainly not *all*) programming languages—for instance, the SK-calculus has no built-in notion of arithmetic. Yet, we say that combinatory logic supports arithmetic, because we can program combinators that take care of all the conventional arithmetic processes.

As it turns out, inter-program reference (reference to the code or results of one program from inside another) is universally supported by all programming languages, since we can program, in any Turing-complete language, a Universal Turing Machine (UTM), which can take as input the code for any other Turing machine, and simulate the running of that machine inside the UTM [217, pp 241-6]. In other words, any one programming language can be used to precisely emulate the workings of any other programming language. So program reference is possible in *all* programming languages. Therefore, there is nothing illegitimate about our introduction of the program-reference

notation `Pk.var[t]`. While it is not a universal primitive feature of all languages, any language can be made to emulate it. The same can be said for the `LOOK()` function, for that matter, which is clearly not a primitive, but which we assume can be implemented in any programming language, given our assumption of Strong AI. We could have avoided having to address this whole issue, of course, if we had actually built these features from scratch using logical combinators (and in the case of program reference, that wouldn't even be all that impractical, although it would still be tedious and completely unnecessary given the formal equivalence of all Turing-complete languages).

7.5.2 The Syntheticity of Interference Effects

A sceptical reader might still remain suspicious that perhaps the kind of interference described above is not "real interference". Since program reference is not inherent to programs, one surely cannot claim, in our example above, that P_3 and P_4 *really* cancel each other out, since one cannot even say that each program is *really* even referring to the other, in the first place.

My response to this is twofold. First, *destructive interference in real-world quantum mechanics is every bit as much an artifact of perspective* as it is in this toy example. Recall that, if state $|\psi\rangle$ has zero amplitude for position x, then there may be one basis where we resolve the state into a superposition of positive and negative amplitude for x, which thereby cancel out; but there may also be another basis where there is no amplitude for x (trivially, the basis where $|\psi\rangle$ is the lone member of the basis set). Thus, whether the quantum state in question is, or is not, a result of destructive interference, is a matter of which basis we choose. Of course, as we have discussed, we do not have a truly free choice of bases in quantum mechanics, due to the fact that we are observers within the system, so there may still be synthetic *a priori* justification for calling the interference "real", but it is still not an analytic reality.

Interference, then, like wavefunction collapse, ultimately has a purely synthetic character. But note that this is in no way some strange quantum mechanical phenomenon. Constructive and destructive interference are purely synthetic for classical waves too. Assume that Liz and Mark are at opposite ends of a room, each holding the opposite end of a length of rope, which lies slack on the floor between them. Now Mark lifts his end and flicks down rapidly, sending a single pulse down the rope to Liz. If he does this repeatedly, he generates a wave with a fixed periodicity. Now Liz does likewise. Clearly, we will see interference between their two waves. Assume that at position x on the rope, at time t, that *if* only Mark were doing anything, *then* the rope would be at its maximum height (+1 unit amplitude). But *if* only Liz were doing something, *then* the rope would be at its minimum (-1 unit amplitude). But because they are both doing something their two effects cancel out and the rope is at its zero position at x.

Note that the character of this description is hypothetical—and in a sense all descriptions of interference effects have this hypothetical character—because the rope never actually achieves a +1 amplitude at x because of Mark's action (and likewise for Liz's). The idea that there is somehow a +1 amplitude summing with -1 is merely an artifact of taking the perspective that Mark and Liz constitute separate forces that need to be separated out for our analysis. It is just as valid to combine them, and not consider them separate forces at all, in which case there is no need to ever postulate a +1 amplitude at x at time t. There never is such an amplitude *in the rope itself*, and *that* is our ontic entity (or perhaps the rope plus Mark plus Liz), at least for the purposes of this example.

So I don't believe that inter-program interference, due to program reference, is any less a legitimate interference effect than real-life wave interference.

7.6 The Syntheticity of Worlds and the Decomposition of Programs

We have taken the position of analyzing our superposition of program(s) into $P_1 - P_5$. But we could just as well consider them a single program, in which case there is no canceling out of P_3 and P_4, it is just that we only need P_1, P_2 and P_5 to join together to make our single program. There will even be ways to divide $P_1 - P_5$ into separate programs that does not involve combinations of these five programs at all.

Imagine, returning to example #4, that a single program (call it $P_{3,4}$) does *part* of what P_3 does (call it P_{3a}) and *part* of what P_4 does (call it P_{4a}). And another program, P'_2, does what P_2 does, plus the part of P_3 that $P_{3,4}$ does *not* do (call it P_{3b}). Now imagine that P'_5 does the job of P_5 plus the part of P_4 that $P_{3,4}$ does not (call it P_{4b}). This would leave us with $P_1, P'_2, P_{3,4}, P'_5$. We will accomplish this alternative decomposition by adding to P_1-P'_5 the necessary inter-program references to maintain the dependencies between the now split-apart sections of once-whole programs.

We will assume in this case that the same continued-continuer relationships hold as before (our choice before was effectively arbitrary, but followed the sequencing of the program steps for simplicity's sake; now, however, we need to retain the exact same relationships we had before, rather than following the program steps, since in reality, we are supposed to be testing pairs of states for the continued-continuer relationship).

Toy Example #7

```
Program P1
Let a=a0+1
LOOK(a) // 1
Let a=a-1
```

```
Program P2\'
Let a=a0+1          // P2
LOOK(a) // 1
Let a=a-2
LOOK(a) // 1
Let a=P34.a[2]+1    // P3b
LOOK(a) // 2

Program P34
Let a=a0+1          // P3a
LOOK(a) // 1
Let a=a0-1          // P4a
LOOK(a) // 1

Program P5'
Let a=a0+4
LOOK(a) // 4
Let a=a-1
LOOK(a) // 3
Let a=P34.a[4]-1     // P4b
LOOK(a) // 2
```

This is, of course, the exact same overall computation as example #4. It has simply been decomposed into a different selection of programs. Note that while some of these program "refer" to each other, in terms of the notation we have used above, it is just as valid to consider this so-called "reference", as I alluded to earlier, to be simply a matter of mutual information content, which is unavoidable in *any* Turing-complete language, regardless of whether there is any built-in "program reference" function.

Our consistent synthetic histories are now as follows:

At moment 1:

P'_5

$P_1 + P'_2 + P_{34}$

At moment 2:

$P_1 + P'_2 + P_{34} \rightarrow death$

$P_1 + P'_2 + P_{34} \rightarrow P_{2'}$

$P_1 + P'_2 + P_{34} \rightarrow P_{34}$

$P'_5 \rightarrow P'_5$

Note that, while we now have extra compute steps in some of our programs, this has not resulted in any extra synthetic moments. In fact, it seems that the synthetic moments are no longer even necessarily in the same order as the compute steps. This may mean that our way of decomposing the system into programs is contrived, but it does not make it analytically incorrect. It *does*, however,

make it (in a sense) synthetically incorrect, in that it could never make sense for Liz to actually view things this way.

Another hint that our program decomposition is overly-contrived is that, if we take the programs as written at face value, it appears that Liz (or more precisely, her universe) must actually compute the contents of her memory from first principles in order for her to simply continue her conscious state (at the end of P_5', for instance). This would seem incredibly wasteful of both computing time and program length. Surely, this will give P_5' a very long program length, meaning a very low probability, compared to the probability we got in our previous analysis. However, this is not actually a problem, after all. Solomonoff probabilities will not be affected, since these programs must be essentially compressed to their smallest possible length, before probabilities can be calculated. Hence, P_5' as written will never be the shortest possible generator of either Liz's 3 state or her 2 state. Since, to calculate the 2 state (coming after the 3 state), we need to ignore the 3 state and effectively compute a result of $P_{3,4}$ from scratch, both mental states will be far more economically computed by sticking to the original decomposition. Hence, even if we choose to express Liz's situation in terms of the $P_1, P_2', P_{3,4}, P_5'$ decomposition, employing Solomonoff probabilities will effectively re-construct the decomposition of programs back into the more synthetically reasonable form $P_1, P_2, P_{3,4}, P_5$. Thus, the principles of program compression in the context of synthetic unity make our probability calculations resistant to artificially contrived program decompositions, even though programs are our countables, and the "contrived" decompositions are analytically just as valid as any other.

7.7 Macrostate Granularity

This elimination of contrived decompositions highlights a more general feature of algorithmic probability that may be potentially useful in addressing the preferred basis problem in quantum mechanics, which is also about trying to justify the use of a valid but synthetically unreasonable analysis (in this case, the preferred measurement basis). Given what we have seen above, it seems reasonable to hypothesize that, if quantum probabilities are algorithmic in nature, it follows that contrived measurement bases will also be effectively washed out in the calculation of Solomonoff probabilities.

However, to construct a toy example that reflects more the structure of the preferred basis problem, there needs to be more at stake here than merely how we decompose the whole situation into programs. Assume that we have already chosen the most efficient (probable) pre-measurement program decomposition (and that it is $P_1 - P_5$ from example #4). Recall that the synthetic histories are:

At moment 1:

$P_1 + P_2 + P_3 + P_4$

At moment 2:

$P_1 + P_2 + P_3 + P_4 \rightarrow death$

$P_1 + P_2 + P_3 + P_4 \rightarrow P_2$

$P_1 + P_2 + P_3 + P_4 \rightarrow P_3 + P_4$

Recall also that the pre-measurement macrostate that Liz is in (at moment 1) is strictly defined (in terms of Solomonoff complexity) as *all* the programs that produce her conscious mental state, *including* the contrived ones, even though we are usually glossing this as the *optimal* program, since the Kolmogorov complexity will predominate in the Solomonoff series. So we are imagining here that we compress all the programs in the macrostate, and identify the macrostate (for all practical purposes) with the minimum-length encoding. As a finite description of Liz and her world, this encoding will necessarily have a certain coarseness of grain. Longer programs may describe her world (supposedly) in more detail, but if such detail has no effect on Liz's mental state, then it will automatically wash out in the Solomonoff calculation. For instance, let's say that we modify P_3 to get P_3':

Toy Example #8

```
...
Program P3'
Let a=a+1
Let x=86
LOOK(a) ==> 1
Let a=a+1
LOOK(a) ==> 2
...
```

Clearly, this produces the same result for Liz as the original P_3. The fact is that $x = 86$ has no affect on her consciousness, so there is no mutual information between the subsystem "Liz" and the subsystem "x":

$$H(Liz : x) = 0 \tag{7.9}$$

The same would be true if the extra information in P_3 were simply extra digits of accuracy in the decimal expansion of a, rather than the existence of a separate variable:

Toy Example #9

```
...
Program P3''
Let a=a+1.0001
LOOK(a) ==> 1
Let a=a+1
LOOK(a) ==> 2
...
```

This program, too, will compress down to P_3, which will predominate in the Solomonoff series. In other words, there is no mutual information between Liz and the fractional part of a.

$$H(Liz : a - \lfloor a \rfloor) = 0 \tag{7.10}$$

This means that the coarseness of grain of the macrostates will be fixed by synthetic unity, just as they were in our discussion of the preferred basis problem in quantum mechanics. Hence, *if* quantum probabilities are synthetic-unitary algorithmic probabilities, Wallace's contention [227] that there is inherent ambiguity in the coarseness of the macrostates, and hence in the branch count, is incorrect. There will, in fact, be a precise branch count, since each macrostate is identified with a particular conscious state. There will also be a precise coarseness of grain for this macrostate, since it will overwhelmingly be identified with its minimum length encoding (which will predominate in calculating the Solomonoff probability). So it would seem that there might, after all, be a well-defined probability based on branch-counting[60].

When we calculate the Solomonoff probabilities for Liz at the first `LOOK()` command, P_3 will predominate over the longer (and lower probability) P_3' and P_3''. So it is not just contrived decompositions that will wash out of the Solomonoff computation, it is also overly-refined ones (although that is really just a kind of contrivance).

Principle 7.5. *The macrostate for a program (*i.e. *for a microstate) that generates a conscious mental state is, therefore, identified with the* coarsest-grained *algorithm for that mental state.*

When the act of measurement occurs, Liz's consciousness bifurcates (ignoring the death state, of course):

$$\begin{aligned} P_1 + P_2 + P_3 + P_4 &\rightarrow P_2\,(a = 1) \\ P_1 + P_2 + P_3 + P_4 &\rightarrow P_3 + P_4\,(a = 2) \end{aligned} \tag{7.11}$$

Note that this changes the coarseness of the macrostate, since we now need a finer-grained macrostate in order to differentiate between microstates that do or do not contain different versions of Liz.

7.8 More Complex Consciousness Patterns

Let's examine more closely now the issue of just how constrained Liz is, in how she breaks down her generator algorithm. It might seem that she could adopt whatever analysis she wished, but this

[60] Wallace would still be correct, however, that branch-counting is irrelevant, because branches are not the ontic entities of an algorithmic system; *programs* are, and as we have established, programs are in no way comparable to branches. However, Wallace still can't get to this from his own standpoint, since to justify counting programs requires an objectivist and metaphysical/generative interpretation of probability, and Wallace cannot go that route given the decision-theoretic framework his proof is based on.

is partly just an artifact of our highly unrealistic assumption that Liz was the presence of a 0 in a single variable. More realistically, Liz will be a pattern (indeed, a very complex one), and there will surely be ways of analyzing this pattern that will not respect the unity of Liz's personal identity. To give an example—only slightly more complicated than the one we have been using—imagine Liz's identity consisted of $a \neq 0$, as before, but in addition, we add the requirement that $b = 8$. Assume that Liz's consciousness is retained only in P_2' and P_{34} (we will add $b = 8$ at the beginning of just these programs, and $b = 7$ at the beginning of the others). Assume further, that in each of these programs, the setting of a to a Liz-friendly value depends on the fact that $b = 8$, so that there is some dependence between the variables, as there would have to be if they were both part of Liz.

Toy Example #10

```
Program P1
Let b=7
Let a=a0+1
LOOK(a) // 1
Let a=a-1

Program P2'
Let b=8
Let a=P34.a[2]+1    [P3b]
LOOK(a) // 2
Let a=b-7           [P2]
LOOK(a) // 1
Let a=a-2
LOOK(a) // 1

Program P34
Let b=8             [P3a]
Let a=b-7
LOOK(a) // 1
Let a=a0-1          [P4a]
LOOK(a) // 1

Program P5'
Let b=8
Let a=a0+4
LOOK(a) // 4
Let a=a-1
LOOK(a) // 3
Let a=P34.a[4]-1    [P4b]
LOOK(a) // 2
```

Now imagine that in splitting the functionality of P_3 across P_2' and $P_{3,4}$, we actually end up splitting Liz apart. Assume that variable a is dealt with *only* by P_2' (call the result P_2''), and b is dealt with *only* by $P_{3,4}$ (call the result $P_{3,4}'$):

Toy Example #11

```
...
Program P2''
Let a=P34'.b[1] - 7   // P3b
LOOK(a) // 1          //
Let a=a+1             //
LOOK(a) // 2          //        [no observer here]    [ Yet,      ]
Let b=8               // P2                           [           ]
Let a=b-7             //                              [ there     ]
LOOK(a) // 1          //                              [           ]
Let a=a-2             //                              [ is an     ]
LOOK(a) // 1          //                              [           ]
                      //                              [ observer  ]
Program P34'          //                              [           ]
Let b=8               // P3a   [no observer here]      [ here.     ]
Let a=a0-1            // P4a
LOOK(a) // 1
...
```

Analytically, this is just another (equally valid) analysis of the same system: once again, it represents the exact same computation. However, synthetically, this second analysis is extremely contrived, as it has separate programs producing separate parts of Liz's current mental state. We assume here that b must be stored in a long-term memory location, as 8, so that Liz may retrieve it at will, so the fact that we used program reference in $P2''$ to retrieve the 8 does not change the fact that it is stored elsewhere—Liz must compute one of her own memories from first principles every time she wants to retrieve it! Of course, this does not affect the Solomonoff probability, for reasons we have already discussed. The most compressed version of Liz will *not* have to re-compute from first principles, and hence our above analysis is synthetically invalid, even though analytically equivalent to the original decomposition. In other words, Liz cannot (with any seriousness) analyze the situation this way—and certainly not if she wants to compute objective probabilities about herself[61]—because it places different parts of her (at the same moment of her synthetic timeline) into *different*, mutually exclusive programs. Note that there is no conscious observer in program P_2'', and neither is there a conscious observer in $P'_{3,4}$. If programs were worlds, there would simply be no Liz, not in any world! However, there *is* clearly still a Liz here: this is the *same* computation, after all.

This leaves us with no way to compute algorithmic probabilities, given this decomposition, since to do so we must count programs, categorized according to synthetic unity. The numerator count must be the number of programs that generate a given mental state, such as Liz's state $(a = 1, b = 8)$. But *this count does not exist* under the $P_1, P_2'', P'_{3,4}, P'_5$ analysis. Do we count $P'_{3,4}$ and/or P_2'' as

[61] We could just as well (and more realistically) talk about an analyst outside the system computing probabilities *about* Liz. Just to be clear: such an analyst is equally bound to obey synthetic unity in calculating *objective* synthetic probabilities about someone else.

generative of the $(a = 1, b = 8)$ state? Since the state that is generated exists partly in the one and partly in the other program, we cannot avoid counting at least *one* of them (for if these were the only programs to produce this state, counting neither would mean counting nothing for a clearly existing state). On the other hand, we cannot count *both* programs (for if these were the only programs to produce this state, we would be counting twice for that which is clearly generated only once, violating the principles of the generative interpretation of probability).

Therefore, while there may be more than one way to decompose the system into individual programs, there are only certain ways that are consistent with synthetic unity, and that can therefore even be used for a generative probability analysis. The above contrived decomposition is possibly analogous to what we would get, in real-world quantum mechanics, if we insisted on using a non-standard preferred basis for the observation of an outcome from a Stern-Gerlach device.

7.9 Unitary Evolution

The next logical step in this research—pushing the premise even further than I have thus far—would be to show that the symmetry that results in unitary evolution in quantum mechanics has an analog in a corresponding symmetry in the toy model. To wit: we wish to demonstrate that Liz's evolution must be modelled by a unitary, and only a unitary, operator.

There is no obvious *a priori* reason to believe that Liz's evolution needs to be unitary. Certainly, the space of all possible operators that *could* operate on a given state of Liz (or a given real-world quantum state, for that matter) include many that are not unitary. However, synthetic unity ensures that there *will* be constraints *of some kind* on allowable operators. In particular, it requires that any operator on a consciousness that is not intended to model collapse must necessarily generate *all* the possible continuers of that consciousness.

Definition 7.6. *An operator that acts on a conscious observer is a "synthetic-unitary operator" if (and only if) it algorithmically generates all possible continuers of that consciousness.*

Principle 7.7. ***The principle of synthetic-unitary evolution:*** any operator that acts on a conscious observer that is not intended to model collapse is necessarily a synthetic-unitary operator.

Note that to falsify this principle, we would need to demonstrate much more than that there exist non-unitary operators that continue Liz's consciousness. One would probably have to falsify ASU itself, since this principle seems to follow from other principles we have already accepted. Unless we are intentially taking the perspective of one of the continuers (we are modelling collapse, in other words), there is no justification for not including all possible continuers (unless, perhaps, we are merely excluding the maverick continuers).

Question. Is "unitary evolution" in quantum mechanics the same thing as "synthetic-unitary evolution"?

One issue here is the simple fact that in its most *a priori* form, ASU requires us to match up arbitrary states and test them for the continued-continuer relationship. This provides the ordering that then generates our synthetic timeline. Thus, there really *is* no dynamics inherent to ASU. Thus, how can we ask what constraints there are on this dynamical evolution?

But this is not exactly the case, since there are clearly constraints on how we match up continued-continuer pairs, this may define an effective evolution operator. In fact, our expectation that ASU will explain cosmic stability implies that this will be the case. If the optimal program for both continued and continuer states is extremely tiny and global in nature, it is extremely likely that the entire synthetic history is going to be dominated by the optimal program, and that we can effectively consider it to be a single-program history. This means that there will be a single program that mirrors the synthetic history, and which may contain many previous "times", as well as later times. This is the external environment. We will have, then, an extended synthetic timeline that will obey the dynamics defined by the program—and since the program is tiny and global, this will manifest at the local level as universal laws.

So it is likely we *will* have some kind of dynamical law. But will it be unitary? Unitary operators ensure the conservation of inner product, and hence of relative amplitudes and (assuming the Born rule) probabilities. They preserve the essential structure of the state, so that we say they are "information preserving" transformations (which is why they are time-symmetric). To show that this conservation is equivalent to synthetic-unitary conservation, we would need to show that the conservation of Liz's conscious personal identity mandates the preservation of the structure of her model.

Note that by "structure", it seems reasonable to assume that we mean something like what is preserved in unitary transformations, although we can only really make loose comparisons here, since we currently have zero knowledge of what the optimal compression of a human consciousness might look like.

We cannot state at this point that 100% of the structure needs to be preserved, since not all of the structure need be essential to Liz's identity. This would be especially true if the wavefunction were a *local* (decompressed) description of Liz, so that some of the data points were mostly about her environment, and not directly about her personal identity, and thus could be safely distorted non-unitarily without resulting in a distortion of Liz's identity. However, recall that our algorithmic measure is, by the very nature of ASU, an inherently *global* measure, creating a description of Liz that is actually smaller than the local description of her brain. Give this, it is difficult to see how

any significant portion of her structure could be irreversibly distorted without distorting *both* her environment *and* her personal identity. To use a loose analogy, the global representation of Liz is something like a hologram: each local piece of it contains information about Liz *and* large swathes of her environment all rolled into one measure. Thus, it seems almost unavoidable that personal-identity-preserving (synthetic-unitary) transformations must necessarily also be structure-preserving and information-preserving.

An objection that can be raised here is that, assuming there is more than one continuer of Liz's current state, then if we simply "throw away" all but one of the branches, we still have a transformation that is consistent with Liz's past identity, but is clearly not structure-preserving (it could only recover *part* of her previous state if the "time clock" were reversed). However, our definition of synthetic-unitary evolution above precludes this kind of "throwing away" (or projection) of Liz's state. This is not arbitrary. The reason is central to the whole algorithmic approach: the only way to justify an *objective* interpretation of Solomonoff probability is to situate Liz in an ensemble of all possible programs that can produce her mental state. Hence, our synthetic-unitary operator must, as stated above, be an enumeration of *all* the programs that preserve Liz's identity. We cannot throw out some of them on a whim. Hence, given the global nature of the optimal representation, it seems unlikely that any non-information-preserving transformation could fit the bill.

A final caveat: we will never, in practice, calculate a total information measure on Liz. We simply do not have the computing resources to do anything like that in the real world, or even anything vaguely close to it. Instead, we will be only (in effect) partially decompressing her representation, so as to achieve a partly phenomenal, partly wave-based representation, so that we have a *local* representation of a phenomenal *subsystem* that can be informationally isolated from the rest of the wave-based representation. This will be the actual context for any predictions for real-world experiments that might be done to attempt to falsify an ASU theory of quantum phenomena.

While a great deal of the general features of quantum mechanics have turned up in our *a priori* system, I think that to go any further along the path to demonstrating unitary evolution, we need a better model of the compression of mental states. Our toy examples thus far have assumed a single bit for an entire mind (clearly, completely incompressible and lacking in many of the properties needed to take our arguments any further). The next chapter will attempt to take this next step, by making a speculative (but well-motivated) educated guess as to the optimal compression algorithm for conscious states.

8 Outline of an *A Priori* Derivation of Quantum Mechanics

Our examples in the previous chapter were called "toy" not merely because they were very small and simplistic, but also because they attempted to derive aspects of quantum mechanics from basic philosophical postulates, resulting in something that was itself *not* quantum mechanics. Nor should it even be considered a kind of idealized quantum mechanics. The point was simply that many features of this *a priori* toy model turn out to be similar to *some* of the features of quantum mechanics, and that *if* these similarities were to hold up as legitimate correspondences with actual quantum theory, then amplitude-counting might be justified as an *a priori*. But without a more precise correspondence with the physical theory, this remains just a hope.

While our toy examples did show why we should expect some kind of superposition principle, and why program-counting could be expected to lead to interference effects of some kind, there is still a great deal of missing structure in our toy "quantum" systems. We still have no derivation of amplitudes in a Hilbert space *as* program counts. In addition, there are other aspects of the basic quantum postulates—especially unitary evolution itself—that remain completely unexplained in our toy examples. So, while these thought experiments are quite illuminating, they take us only a very small step towards a full *a priori* justification of the Born rule.

In this chapter, while I cannot claim to take us all the rest of the way in a fully rigorous and completely justified way, I would like to try to complete the derivation in outline, at least. There may be gaps and weak points, but the result will provide the basic framework to build on for a more rigorous and complete development, and I hope that you will see the gaps as well-motivated and not completely *ad hoc*. This is the latest refinement in an ongoing project to rationally reconstruct the postulates of quantum theory [169, 172, 174].

My main move in this chapter will be an assumption about the actual compression algorithm the universe uses to "compress" a conscious state into the shortest possible program.[62] While this will

[62] The universe does not actually perform this compression, as an algorithm like this, hence the scare-quotes around "compress". Rather, the compression algorithm is a means by which we can imagine computing the shortest (and hence most probable) program that generates the universe. In reality, we cannot even say we *are* the most probable, but only "average" or "typical", so even this way of talking is approximate. Remember, our system relies on self-selection, so all that is really happening is that we are a random pick from programs that generate the universe—it is just a helpful turn of phrase to say that the universe actually "compresses" our consciousness (not so different in kind from

deliver most of the rest of the analytic structure of quantum mechanics, the role of amplitude here will come about differently than for its analogue in the toy examples, and (at least at first glance) we will appear to still be faced with a new and more vigorous version of the Born rule objection. I will argue, however, that this does not hold up on closer inspection, and that we can expect the Born rule to hold, not due to an Everett-style branch-counting frequentist argument, but due to the result of Gleason (see §3.3.12). I will argue that the arguments against Gleason as an MWI Born rule proof do not apply under the assumptions I will make in this chapter (although these assumptions can certainly be contested).

8.1 Technological Data Compressors

One general feature of quantum mechanics that was missing in our toy examples was its periodic or wave-like structure (although it is notable that we still had an interference effect, even without assuming waves or periodicity). If we consider even a modestly more realistic model of Liz's consciousness—one, at least, that admits of significant data compression—then an argument can be made that the inter-program interference I have described would take on a more wave-like character, like that in quantum mechanics. The simplest argument for this comes from the straightforward observation that *(1)* compression requires the discovery of redundancy or symmetry in the dataset, and *(2)* periodicity is one of the basic, and most common, kinds of symmetry.

However, to take this line of reasoning any further, we will need to postulate some kind of actual compression algorithm for human conscious states. The analysis that would be required here to *prove* a particular algorithm to be the optimal one might very well be massively complex, requiring a complete understanding of human psychology, and perhaps even *(a priori)* knowledge of the fields of cosmology and evolutionary biology.[63] However, it is possible to make a quite educated guess about the basic structure of such an algorithm, based on the current state of the art of computer compression technology, and the kind of compression we might expect to be required for human consciousness, based on current *a posteriori* experience.

Real-world data compressors come in two basic flavours:

1. **Lossy compression** allows some ambiguity in the actual bit-sequence that must be reconstructed from the compressed data, while
2. **Loss-less compression** requires that the *exact* bit-sequence be recoverable.

---

asking things like "If God picks a world from a hat, what is the most likely kind of world he will pick?").

[63] I am not claiming that complete knowledge of *all* these fields would really be required to derive the correct compression algorithm. I have no idea, for instance, how much psychology versus cosmology might actually be needed. All I am claiming here is that we cannot yet say how much knowledge of these fields would be required, but that it seems likely more is needed than an analytic understanding of the five quantum postulates.

In other words, lossless compression is required when there is a precise, known analytic relationship between the original and recovered data, while lossy compression is used when the criteria for the recovered data are synthetic, and dependent on human perception. Facts about human perception are thus relevant to the design of lossy compression algorithms.

Compressing a human consciousness would, I believe, clearly require a lossy compression algorithm. This is especially so under the Strong AI view of consciousness, where multiple different realizations of the same algorithm can be said to compute the same conscious state. In this case, it is clear that there is a great deal of latitude in the actual bit sequence that we recover from the compressed representation. It would be difficult to justify requiring that such a compression be loss-less.

Our compression algorithm should take a form closer, then, to that of *lossy* compressions algorithms, such as those used for images and video. These kinds of data use lossy algorithms precisely because it is very difficult to state precisely and analytically what the requirements are for adequate reconstruction of the original. Given a photograph, for instance, if we look at it on the level of very small clusters of pixels, the precise local configuration of these pixels could feasibly be modified (even quite radically) throughout the entire image, and we could still end up with essentially the same effect on the eyes of the perceiver. This places the compression of images and movies inherently in the lossy domain. Human consciousness itself will, I argue, have these same basic qualities, and for very similar reasons.

8.2 Analogues to the Quantum Postulates

In this section, I will develop a lossy-compression version of ASU, based on what is generally known about technological data compressors. I will frame my development of these ideas as an outline (not a thorough and complete) synthetic *a priori* derivation of the five quantum postulates.

8.2.1 Synthetic-Unitary Representation [Quantum Analogue: Postulate #1]

If we look at most proven-effective[64] real-world lossy data compression algorithms, we find that they work by the same basic scheme: by transforming the original N data points $\{\langle x \mid \psi\rangle : x = 0 \cdots N-1\}$ into R frequency amplitudes $\{\langle k \mid \psi\rangle : k = 0 \cdots R-1\}$, by way of something like a discrete Fourier

[64] "Proven" here, of course, does not mean *analytically* or even *a priori* proven, but rather it simply means that the algorithm has been shown, *a posteriori,* to be very efficient in real-world use.

transform (DFT):

$$\langle k \mid \psi \rangle = \sum_x \langle k \mid x \rangle \langle x \mid \psi \rangle \qquad \text{or} \qquad \psi(k) = \sum_{x=0}^{N-1} \psi(x) e^{-ik\frac{x}{N}\tau} \tag{8.1}$$

$$\langle x \mid \psi \rangle = \sum_k \langle x \mid k \rangle \langle k \mid \psi \rangle \qquad \text{or} \qquad \psi(x) = \frac{1}{N} \sum_{k=0}^{R-1} \psi(k) e^{i\frac{k}{R}x\tau} \tag{8.2}$$

where $R \ll N$.

The amplitudes are complex, as is the original data, in general. However, in practice the original data is usually real-valued (and, for our purposes, there is no reason for it not to be). The inverse transform (or reverse DFT) allows us to transform the data back into the localized non-frequency domain, and thereby to decompress and recover the original data.

Definition 8.1. The set of amplitudes resulting from a DFT compression from resolution N to R is returned by the $\mathcal{F}()$ function:

$$\mathcal{F}\left(\{\langle x \mid \psi \rangle : x = 0, \ldots, N-1\}, R\right) = \{\langle k \mid \psi \rangle : k = 0, \ldots, R-1\} \tag{8.3}$$

with the inverse DFT given by $\mathcal{F}^{-1}()$ returning the decompressed reconstruction of the original data:

$$\mathcal{F}^{-1}\left(\{\langle k \mid \psi \rangle : k = 0, \ldots, R-1\}, N\right) = \{\langle x \mid \psi \rangle : x = 0, \ldots, N-1\} \tag{8.4}$$

Given that this basic scheme is almost ubiquitous in lossy compression technologies, no matter the application, it seems that we can now put forward a very well-motivated hypothesis about the most likely compression algorithm applicable to real-world synthetic unity.

Assumption 8.2. ***The Weak DFT Hypothesis****: the compression algorithm that produces the optimal encoding of a typical conscious state m is a Fourier-based transform compression algorithm, by which we will mean that it is, in its essential aspects, basically a discrete Fourier transform, where the high frequencies are truncated at the lowest possible point that still encodes the original conscious state, resulting in a list of R_m complex frequency amplitudes a_k for $k = 1 \cdots R_m$, each of bit-length r_k. It is allowed that there may still be further symmetries remaining in the data after such a transform is performed.*

The astute reader will already see, recalling §2.4, that this will take us much further than our previous toy examples, and will provide us with most, if not all, of the remaining analytic structure of quantum mechanics, since the Fourier transform above *is* the discrete form of the solved version of Schrödinger's equation. This gives our informal toy superposition principle a precise and mathematical foundation, that is happily exactly the same as that of quantum theory, but with a specific connection to ASU, to aid us in deducing a probability rule.

The weak DFT only assumes that our consciousness-compressor is Fourier-*based*. It presents the discrete Fourier transform as a framework in which to construct the optimal compression, but does not assume there are no further symmetries that will need to be compressed out. However, for most of what follows, we will use the stronger assumption that it *is* a strict DFT that is required, with no further symmetries remaining in the data:

Assumption 8.3. ***The Strong DFT Hypothesis:*** *the optimal compression of a typical conscious state m is a strict DFT transform from a locally optimal representation to $|\psi_m\rangle$ in frequency basis $\{|k\rangle\}$ of dimension R_m,*

$$\langle k \mid \psi_m \rangle = \sum_x \langle k \mid x, t_i \rangle \langle x, t_i \mid \psi \rangle \tag{8.5}$$

where the amplitudes $\{\langle k \mid \psi_m \rangle : k = 0 \cdots R_m - 1\}$ are represented with fixed-precision bit-lengths, and contain no further symmetries. We assume that this implies an arrow of time of successive mental states, indexed by t, that each continue the previous mental state (the compression in the equation above is performed for $t = t_i$, which is taken to be our starting time, the time of the original mental state m). ***[Quantum Analogue: Postulate #1, Superposition]***

Unitary transforms can then be performed from the frequency basis to any arbitrary basis. Of particular interest, however, is the inverse transform (decompression) back to the local representation in basis $\{|x, t\rangle : x = 0 \cdots N_t\}$ for any other arbitrary time t (including the original time t_i):

$$\langle x, t \mid \psi_m \rangle = \sum_k \langle x, t \mid k \rangle \langle k \mid \psi \rangle \tag{8.6}$$

and the transform from one synthetic time t_1 to another time t_2,

$$\langle x, t_2 \mid \psi_m \rangle = \sum_x \langle x, t_2 \mid x, t_1 \rangle \langle x, t_1 \mid \psi \rangle \tag{8.7}$$

where we order our times starting at t_0 and according to the dimension of the local basis,

$$t_1 < t_2 \Leftrightarrow N_{t_1} < N_{t_2} \tag{8.8}$$

Thus, $|\psi_m\rangle$ takes an increasing number of bits to describe locally, as t increases. This is *not* an entropy increase, however, since these are all unitary transforms, which preserve information and have equal entropy, $H(m)$.

This stronger version of the DFT hypothesis is not really justifiable *a priori* over the weak version, but it is the logical scientific hypothesis to put forward, given that we do not currently know what "remaining symmetries" might exist in human consciousness—hence, the simplest approach is to begin with the stronger assumption, since it is analytically simpler, and look to add in further symmetries at a later date, as more becomes known. It is always possible that the symmetries really are strictly periodic, in which case there may be no need to delve very deeply into the details of the

workings of human consciousness in order to understand nature (or, at least, in order to understand quantum mechanics), but this is a question that I will leave for another time.[65]

The forward and inverse DFTs are simply taking the same complex vector and translating back and forth between different orthonormal bases. Note, however, that while the DFT and inverse DFT transformations are *in general* symmetrical, and translate between representations at the same resolution, here we are using them as compression/decompression algorithms, and so the resolution of the frequency domain will be much lower than that of the local domain. This asymmetry is crucial to the application of the DFT to algorithmic synthetic unity.

Of course, the continuous versions of these transforms,

$$\langle k \mid \psi \rangle = \int_{-\infty}^{\infty} \langle k \mid x \rangle \langle x \mid \psi \rangle \, dx \tag{8.9}$$

$$\langle x \mid \psi \rangle = \int_{-\infty}^{\infty} \langle x \mid k \rangle \langle k \mid \psi \rangle \, dk \tag{8.10}$$

cannot be used in this way, since the resolutions of both are infinite, and there can thus be no compression.

8.2.2 Synthetic-Unitary Evolution [Quantum Analogue: Postulate #2]

While the entropy of $|\psi_m\rangle$ remains the same in whatever basis we choose, the fact that there is an increase in *local* bits (or the dimensionality of the local basis) does seem to suggest the appearance of some kind of entropy increase from the perspective of the observer. It also means that we can define a minimum time t_0 such that the number of bits can simply be taken to be the same as the global representation. Hence, the global optimal compression in basis $\{|k\rangle\}$ is identical to the so-called "local" representation at $t = t_0$,

$$R_m = N_{t_0} \tag{8.11}$$

This yields a conception very much like an optimally compressed program serving as the "initial state" of the universe at t_0, which then evolves over time unitarily. Each subsequent time sees an increase in dimensionality of the basis, which hence becomes gradually more and more local and less and less global:

$$N_{t_k} < N_{t_{k+1}} \tag{8.12}$$

[65] Although it seems, in fact, to be the case, since Schrödinger's equation is simply a Fourier transform. By "quantum mechanics", I mean the five postulates. There is still, of course, the possibility that the details of particle physics and general relativity, if further developed along ASU-DFT lines, might require further details about consciousness. However, this question is beyond the scope of this dissertation.

Since the transformations through synthetic time are unitary, this yields a more precise statement of our principle of synthetic-unitary evolution from our toy examples.

Principle 8.4. ***Synthetic-Unitary Evolution****: Any transformation of a complete wavefunction (the entire wavefunction of an observer's universe), that does not involve perspectival collapse (and hence entropy increase and information loss), must preserve the identity of the observer, and must therefore be a symmetry transformation on the optimally compressed representation (by the DFT Hypothesis), which by* [232] *must be either a unitary or antiunitary operator. Since the optimal compression is a DFT of the local data, evolution through synthetic time is representable as a (unitary) DFT from one basis to another.* ***[Quantum Analogue: Postulate #2, Unitary Evolution]***

If we wish to ask what the dynamics of the phenomenal world (*i.e.*, the universe described in the *local* basis space) ought to be, it makes some sense to go ahead and work with the continuous transform, as we will be working in a basis with much higher than optimal dimensionality anyway, and probabilities are derived from the optimal basis, not the local basis:

$$\langle x,t \mid \psi \rangle = \int_{-\infty}^{\infty} \langle x,t \mid k \rangle \langle k \mid \psi \rangle \, dk \tag{8.13}$$

To express this in terms of local dynamics for the perceiver, we need to express it as a differential in terms of extended synthetic time. Recall from §2.4, that the above equation can be interpreted as a superposition of plane waves, in terms of space and time, and that under that interpretation, the above inverse Fourier is a solution to the time-dependent differential equation:

$$i\frac{\partial}{\partial t} |\psi\rangle = \hat{H} |\psi\rangle \tag{8.14}$$

where $\hat{H}$ is a unitary operator. This is, of course, Schrödinger's familiar dynamical equation, interpreted here as the natural form for the dynamics of a phenomenal world with an arrow of time, in an algorithmic synthetic-unitary system where the optimal compression algorithm is assumed to be the same as the most general and widely successful technological solutions used today in lossy applications.

While this does not *prove* that the DFT is the optimal compression of conscious states, it clearly lends considerable empirical support to the idea. And while not all lossy compression schemes are exactly DFTs, the vast majority are "Fourier-based": they are either DFTs or can be described as DFTs modified to take into account some further (non-periodic) symmetries in the data. One of the more common examples is the "discrete cosine transform" or DCT (used in the JPEG and MPEG algorithms) which is used with data that displays *even* symmetry, in addition to periodic symmetry[66]. The "discrete sine transform" or DST can be used to take advantage of *odd* symmetry.

[66] An example of even symmetry would be a human face—essentially mirrored left to right—or the function $f(x) = x^2$. An example of an odd symmetry would be the Jack of Spades, or the function $f(x) = x^3$.

In fact, for any data set that displays some kind of non-periodic symmetry—in addition to periodic symmetry—we can use a Fourier transform with some kind of further modification designed to take advantage of the non-periodic symmetry. Most practical data compressors, in fact, are not pure DFTs. Usually, when one performs a DFT compression, there will still be remaining symmetries in the resulting amplitude list, and most practical transform compressors add a grab-bag of other techniques to remove these further symmetries.

I do not intend, therefore, to make any dogmatic claims that the optimal compression of a human consciousness is *exactly* a DFT. The larger intent here is only to assume it to be a Fourier-*based* algorithm, as it seems entirely plausible that there could be non-periodic symmetries inherent to human consciousness, in addition to periodic ones. However, given the almost ubiquitous presence of Fourier and Fourier-based transforms in lossy compression technologies, it seems a very fair bet that whatever further modifications might be needed, the basic algorithm for compression of a human consciousness will still be Fourier-based. And since we do not know enough at this point to make a very well-educated guess as to what the non-periodic symmetries might be (if any), the simplest assumption for the time being is to assume the basic DFT algorithm, and see how close to real-world quantum mechanics this might get us. While not entirely justified, I think this constitutes a very reasonable *a priori* educated guess.

And—seeing that the general solved equation for the mechanics of the universe *is* a Fourier transform, according to a great deal of synthetic *a posteriori* evidence—this new hypothesis takes us most of the rest of the way to quantum mechanics proper. Of course, this does not make the toy examples in Ch. 7 worthless, as they are operating on a fundamentally more *a priori* level, not relying on any assumptions at all about the exact nature of the consciousness-compression algorithm. The DFT hypothesis, although I think it is well-motivated, is a very different sort of assumption than the rationalist philosophical assumptions used in the toy examples. The DFT hypothesis is more like a shrewd preliminary guess, based on an observation of the current state-of-the-art of computer technology—a largely *a posteriori* analytic justification for what is supposed to be, in principle, an *a priori* synthetic claim.

However, the *a posteriori* analytic evidence *is* quite strong, given how common the use of Fourier-based algorithms are in lossy compression—so much so, that it is perhaps difficult to credit making any other hypothesis, given our current state of knowledge of consciousness. In fact, I would suggest that if the technological problem of compressing a human consciousness were presented to just about any qualified engineer—even if we could erase from his or her mind any knowledge of quantum theory—the almost certain response would be "try a Fourier transform". On the other hand, it is important to note that the DFT is *also* very widespread throughout all kinds of data analysis, in

general. The DFT is one of the most widely applied bits of mathematics. It has even been said (although I have no idea if this is based on hard data or is just a popular perception) that over 90% of all computation (as measured in CPU cycles) occurring in the world at any given moment is dedicated to the computation of discrete Fourier transforms, so important and ubiquitous are they. The DFT is commonplace throughout all manner of mathematical analyses, especially digital signal processing, and this actually *lessens* its value as a falsifiable prediction for ASU. To see why, consider that Popper [158] did not consider all falsifiable statements to be on an equal footing. Some falsifiable hypotheses, he said, take a greater risk, meaning they are less likely, assuming that our hypothesis is *false*. Thus, when these hypotheses pass their falsification tests, they are more strongly corroborated than falsification tests that take less risk. Of course "less risky" here is a purely informal heuristic, and Popper neither gave, nor intended, a precise definition for it—nor can I think of an appropriate one. But imagine that our considerations about the workings of human consciousness and data compression had lead us to hypothesize some very esoteric consciousness-compression algorithm, one that has no known uses outside of compression, and is not even that widespread as a compression algorithm (perhaps it is even a brand new algorithm, never before used, that relies on very precise details about human perception or even human consciousness itself). Such an algorithm, as an hypothesis, takes great Popperian risk, as it is highly *likely* given that our hypothesis is *true*, while being highly *unlikely* given that the hypothesis is *false*. If *this* algorithm had turned out to have the same analytic structure as the solved Schrödinger's equation, it would be a truly stunning corroboration of algorithmic synthetic unity.

The DFT, however, has so many uses in such a stunningly wide variety of applications, unrelated to quantum mechanics, that Gilbert Strang, who said that "the importance of Fourier transforms is almost unlimited," called the fast algorithm for computing the DFT "the most valuable numerical algorithm in our lifetime." [209, p.495] So even if ASU is *false,* it does not seem that something as commonplace as a Fourier transform should be so surprising to find at the heart of quantum theory, as we can imagine that there may be many other hypotheses which might *also* be centered around a Fourier transform. On the other hand, the equation still *does* corroborate algorithmic synthetic unity, however weak or strong we choose to interpret that corroboration to be.

8.2.3 Synthetic-Unitary Branching [Quantum Analogue: Postulate #4]

By saying that the original non-frequency-based representation at time t is "locally optimal", I simply mean that it is already "compressed" within this narrow context. In other words, it is something like a zipped computer file of the complete neurological state of a brain. According to ASU, its size will still be much greater than the dimensionality of the optimal frequency basis. Since a transform from the

optimal t_0 to the later original time t_i reproduces the original locally optimal brain state, it follows that if we transform successively from t_0 to t_i, we will only achieve a fully localized representation when we are back to t_i, our starting point. Further transforms to later times will produce larger localized representations, but will still preserve information. Hence, we can expect, in general, to have observer branching after time t_i, since the optimal representation is presumably consistent with more than one future for the consciousness it describes, as it is presumably underspecified for later times, and since the transformations are always unitary, the information about multiple future observers—all consistent with m—cannot be excised from the wavefunction during time evolution, else the evolution would not be unitary.

Once we take into account collapse, we are excluding certain particular observers, we are breaking the symmetry of the DFT and it will now take extra information to distinguish which path through the branching *this* particular observer took. This means that information content, and hence, entropy will increase. Since the branching only makes sense from the local perspective of the observer, this is a more rigorous expression of the "apparent local entropy increase" we spoke of earlier. We are essentially removing all but one of the possible future observers from the local representation, giving us a new starting point m', with a new optimal compression at t_0. While we are *removing* bits locally (by throwing out most of the observers), this requires *adding* bits to the optimal global representation. Both of these changes thus represent an *increase* in entropy. Think of the added bits as the extra information needed to specify which branching the observer will take after time t_i.

Assume we start with a conscious mental state m, whose optimal encoding is $|\psi_m\rangle$ in basis $\{|k\rangle\}$. The entropy of the wavefunction after the measurement operation, but before collapse/branching[67] is the same as the pre-measurement entropy (since unitary transformations without collapse do not change our information measure) and equal to the dimension of the frequency basis:

$$\begin{aligned} H(\psi_m) &= H\left(\psi'_m\right) \\ R_m &= R'_m \end{aligned} \tag{8.15}$$

where $|\psi'_m\rangle$ is the (fictitious) state of the system "after" measurement, but before branching.

The post-measurement entropy (after branching) is

$$H\left(\psi_{m'}\right) = R_{m'} \tag{8.16}$$

where m' is a (selected) conscious state after branching, and $|\psi_{m'}\rangle$ the corresponding state of the wavefunction. The entire measurement process can be summarized as

$$|\psi_m\rangle \Longrightarrow |\psi'_m\rangle \rightarrowtail |\psi_{m'}\rangle \tag{8.17}$$

[67] But for the same reasons as in the standard Everett interpretation, the idea of there being an actual state that exists after a unitary transformation, but before branching, is really just a convenient fiction.

Branching, however, increases entropy[68]—but in that case, m' above is really a superposition of two mutually exclusive observers—call them A and B—of which only one will be (synthetically) realized:

$$|\psi_m\rangle \Longrightarrow |\psi'_m\rangle \overset{A}{\rightarrowtail} \left|\psi_{m'_A}\right\rangle \tag{8.18}$$

For simplicity, we may use R_A for R_{m_A}, and $p(A)$ for $p(m_A)$, so long as no confusion results.

World-counting would make m_A and m_B equiprobable, so that $p(A) = p(B)$, but clearly this cannot generally the case, since $|\psi_A\rangle$ and $|\psi_B\rangle$ are macrostates, not microstates. The question, however, is whether we get the Born rule here, or something else entirely.

Clearly, entropy will increase with respect to the branches,

$$R_A, R_B > R_m \tag{8.19}$$

but not the non-collapsed, unitary system,

$$R_{m'} = R_m \tag{8.20}$$

where each mental state m has Solomonoff probability,

$$p(m) = b^{-R_m} \tag{8.21}$$

where b is the numerical base for our information measure.

Since A and B are mutually exclusive and exhaustive of the possibilities, it might seem that we should have, by additivity and summation to unity:

$$\begin{aligned} p(m) = p(m') &= p(m_A) + p(m_B) \\ &= b^{-H(A)} + b^{-H(B)} \\ &= b^{-R_A} + b^{-R_B} = 1 \end{aligned} \tag{8.22}$$

However, this is not the case. While A and B are mutually exclusive *events*, they are not *informationally independent*. In other words:

$$H(A : B) \neq 0 \tag{8.23}$$

In fact, they are not even approximately independent. Outcomes A and B are fully described by m_A and m_B, which are only *slightly* different versions of the same person/universe, and so (on a cosmic scale, at least) share almost *all* of their information! However, A and B are clearly mutually exclusive *as mental states*, and so represent mutually exclusive events, so we certainly should be able to invoke additivity and summation to unity for probabilities on these events. But, while we

[68] This does not exclude the possibility that one could define some other entropy measure that increases without requiring a branching.

can add probabilities, we cannot thereby add the information-based probability formulae (b^{-H}) as we tried to do above in (8.22).

Since m_A and m_B differ by only a small number of bits, from this global perspective, the non-unitary branching process could still be characterized as *almost* a symmetry transformation. It is *almost* information-preserving—so much so, that we may tend to think of it as the *addition* of a few bits of randomness to an otherwise identical universe, rather than as an actual *loss* of structure. However, addition of random bits and loss of structure are the same thing in this case. We *lose* the structure of pruned branches, the result of *adding* random bits to the optimally compressed state.

We can quantify the information loss in algorithmic terms, since

$$H(m) = H(m') = H(A:B) = H(A) + H(B) - H(A,B) \tag{8.24}$$

The mutual or shared information $H(A:B)$ between the two branched mental states m_A and m_B is just the information in the original, coherent state m or m', which can be viewed as a superposition of m_A and m_B, in the right basis.

$$\begin{aligned} H(A:B) &= H(m) = H(m') \\ &= H\left(a_A \left|\psi_A\right\rangle + a_B \left|\psi_B\right\rangle\right) \end{aligned} \tag{8.25}$$

The change in entropy within branch A, due to the branching, is therefore the conditional entropy

$$\begin{aligned} \triangle H_A \quad = H(A|B) \quad &= H(A) - H(A:B) \\ &= H(A) - H(m) \end{aligned} \tag{8.26}$$

We can now express a more precise version of our principle of synthetic branching of histories.

Principle 8.5. ***Principle of Synthetic-Unitary Branching:*** *Given mental state m, at time t_i, whose optimal compression is a DFT from a local representation to $\left|\psi_m\right\rangle$ in frequency basis $\{\left|k\right\rangle\}$, we expect in general that observers will experience a synthetic collapse to one observer m' out of all the observers in an effective branching of observers, for which there must be at least one "preferred basis" $\{\left|o_i\right\rangle\}$ in which the wavefunction is a sum of terms, each one of which represents one possible outcome:*

$$\left|\psi_m\right\rangle = \sum_i c_i \left|o_i\right\rangle \tag{8.27}$$

assuming, as seems reasonable, that it is not possible (synthetically) to branch an observer in multiple possible ways. After an observation, only one of the members of the preferred basis will remain, from the perspective of any one particular observer,

$$\sum_i c_i \left|o_i\right\rangle \rightarrowtail \left|o_A\right\rangle \tag{8.28}$$

by definition of the preferred basis. ***[Quantum Analogue: Postulate #4, Collapse]***

The change in entropy is *not* a measure of the number of bits that distinguishes outcomes in the position (x) basis. It is measure of difference in the (global) momentum (k) basis. As we have expressed things above, this will result in a greater R value:

$$\triangle H_A \quad = \quad R_A - R_m \tag{8.29}$$

This delta entropy should represent the resulting probability of the phenomenal outcome. Note that $\triangle H_A$ is not necessarily the same as $\triangle H_{\neg A}$, since A and $\neg A$ themselves do not necessarily have the same entropy. Thus, the probability of an actual outcome is the probability of the resulting mental state *given* all the information that is contained in the alternative outcome(s). The demands of summation to unity can be dealt with by a normalization constant, call it Z:

$$\begin{aligned} p(A) &= \frac{1}{Z} b^{-H(A|B)} = \frac{1}{Z} b^{-\triangle H_A} = \frac{1}{Z} b^{-(R_A - R_m)} \\ p(B) &= \frac{1}{Z} b^{-H(B|A)} = \frac{1}{Z} b^{-\triangle H_B} = \frac{1}{Z} b^{-(R_B - R_m)} \end{aligned} \tag{8.30}$$

The normalization constant is simply the sum of the probabilities:

$$Z = b^{-(R_A - R_m)} + b^{-(R_B - R_m)} \tag{8.31}$$

When A and B are exhaustive of all possibilities, B is simply the complement of A, call it $\neg$A, and this simplifies to

$$\begin{aligned} p(A) &= \frac{b^{R_{\neg A}}}{b^{R_A} + b^{R_{\neg A}}} = \frac{1}{Z} b^{R_{\neg A}} \\ p(\neg A) &= 1 - p(A) = \frac{b^{R_A}}{b^{R_A} + b^{R_{\neg A}}} = \frac{1}{Z} b^{R_A} \end{aligned} \tag{8.32}$$

So, generally,

$$\begin{aligned} p(x) &\propto b^{-R_x} \\ &\propto b^{R_{\neg x}} \end{aligned} \tag{8.33}$$

In other words, the probability of an outcome is proportional to the power set of the optimal dimension of the alternative outcome(s). Or, equivalently, it is inversely proportional to the power set of its own optimal dimension (assuming normalization in both cases). This is identical to taking probability as proportional to the change in entropy.

This all looks superficially like the same form that we rejected in (8.22), where we tried counting the bits for each outcome in the local representation. However, what we have here is actually quite different. R_A above is the *optimal* dimension for an outcome, in the global basis. Since information is preserved under unitary transforms, this information must be extractable from the

optimal representation of m, but it is not a count of bits in that representation. It is a count of bits in two alternative optimal representations of the two outcomes that are encoded in m. So the question of our probability rule becomes the question of how the above probability is encoded in the amplitudes of the original m, when transformed from the optimal into the preferred basis.

8.2.4 Synthetic-Unitary Probability [Quantum Analogue: Postulate #5]

It might seem, in line with the Born rule objection, that the outcomes in ASU-DFT should all be equally probable. After all, the wavefunction has the same informational structure as the compressed version, and if probability is informational and conserved, it would seem that each amplitude (under strong DFT) takes up the same number of bits, and so should have the same probability. This is the essence of the Born rule objection, expressed in ASU-DFT terms, and we are back to asking, along with the Born rule objectors: whence comes amplitude dependence? Here it seems that probability should vary with the bit-lengths across outcomes, *not* with amplitudes. Given the strong DFT hypothesis, the amplitudes are fixed-precision and all take up the same storage space, so they have equal bit-lengths. Hence, algorithmically, they have equal probabilities.

However, this argument falls (once again) into the trap of thinking of these alternative outcomes as if they were independent information carriers. They are not; they are potentially massively interdependent. They can be mutually independent *as* outcomes, since it is not possible for one particular observer to experience both simultaneously, while still being mutually interdependent *as programs*. This is the core reason why it is possible to have interference between mutually exclusive possibilities, in defiance of classical wisdom.

Assume (without loss of generality) that m_A and m_B are optimally described with differently-sized programs:

$$H(A) \neq H(B) \tag{8.34}$$

These alternative outcomes are optimally compressed in the original DFT, from the perspective of observer m. This is not the same as an optimal compression of the simple juxtaposition of m_A and m_B, call it (m_A, m_B). It would seem, *prima facie*, that juxtaposition has discarded amplitude structure. It is not clear, however, that the amplitude structure could not be somehow reconstructed from the juxtaposition. This is an important question: can we reconstruct m given (m_A, m_B)? On the one hand, we can't do so straightforwardly, since the juxtaposition does not tell us the amplitudes. But that does not necessarily mean that we can't work backward from the juxtaposition to m, from which they branched, especially given that m_A and m_B have high mutual information.

It would seem that at least some information *must* be thrown out, however. Else, there would be no entropy increase, and we would have

$$H(A, B) = H(A : B) \tag{8.35}$$

which is impossible unless m_A and m_B are actually identical, which might be nontrivially possible in a universe where we could make exact copies of mental states (rather unlikely, but in any case, we have already assumed the general case where A and B have different optimal bit-lengths, so they cannot be identical or even have identical entropies).

However, it is perfectly possible for information to be lost *analytically*—which is what matters for probability counts—but not synthetically. For instance, even if we cannot reconstruct the entire amplitude structure, analytically, it is possible that we *could* do so, taking into account the synthetic information that m_A and m_B are both the only possible continuers of m. The synthetic context might provide the missing information (in the language of the Sleeping Beauty debate, the missing information is *centred* information). Nonetheless, the fact still remains that, analytically,

$$H(A, B) > H(A : B) \tag{8.36}$$

No matter how much or little of the amplitude structure of m remains encoded in (m_A, m_B), it would seem that it could not *all* be lost. In particular the information contents of the various outcomes, and thus their probabilities, cannot have been lost, since we can simply calculate these from the bit-lengths of m_A and m_B.

A Born rule objector might ask us to count the bits needed to represent c_A, and use that to compute the probability, as $b^{-L(c_A)}$. This would result in equi-probability, since amplitudes all have equal bit-lengths. But, as mentioned earlier, it doesn't follow that we count bits. In fact, it follows from ASU-DFT that we do *not*. However probabilities *are* encoded into $|\psi_m\rangle$, since it is an optimal compression, but the difference in bit-lengths that generated the different probabilities in the first place will be equalized in the optimal compression, by the strong DFT hypothesis. The list of amplitudes is an algorithmically random list of numbers, with all the pre-existing symmetry removed. Hence, the probabilities will be encoded strictly in the *values* of the amplitudes, not their bit-lengths. Note that this statement holds for the optimal compression, not for its representation in other bases. If we perform a DFT to our current synthetic time, we now have a *longer* bit-length representation, for which we cannot say on the face of it that probabilities are independent of bit-lengths. However, we *can* say that probabilities will be conserved across the DFT, by the nature of probabilities (summation to unity and additivity) and by the noncontextuality of ASU. We also know that both information content and norm-squared amplitudes are *also* conserved across

all DFTs. This makes the Born rule a likely candidate measure, while branch-counting is ruled out (although it is still not a proof).

Theorem 8.6. ***Noncontextuality:*** *given an optimal compression of m in basis $\{|k\rangle : k = 1 \cdots R\}$, and DFTs to two bases $\{|x\rangle : x = 1 \cdots N_x\}$ and $\{|y\rangle : y = 1 \cdots N_y\}$, $N_x \geq R \leq N_y$, if m_A is a continuer of m, then the measure of m_A is a function solely of m_A, regardless of which basis it is considered to be a member of.*

Proof. This actually follows fairly straightforwardly from ASU principles. The probability of m_A can clearly be calculated (at least in principle, it can be limit-computed) using a brute force method from basic ASU principles, independent of measurement context (including whether any measurement or observation is even present).

Since we are still working with complete descriptions here (no observer/environment separation has yet been imposed), we have complete information about the universe in $|\psi_m\rangle$, the wavefunction for m. First, we simply limit-compute the optimal compression of m. Then we perform the symmetry transform that generates all continuers of m. We will need a computable or limit-computable function that can test for continuers, of course, and we will simply assume that such exists without further argument. Since we are insisting that *all* continuers be included, this step can be done without throwing out any information, and there must therefore exist such a symmetry transform. Note that the continuers are still encoded in m. Performing the continuer-DFT transform does not really "generate" them, in the sense of creating new information. It simply re-expresses $|\psi_m\rangle$ in terms where the division between continuers is apparent, so that, in the preferred basis, each continuer corresponds to a separate term $|k\rangle$ with amplitude c_k.

Now that we have thus generated the continuers, we compute the optimal compression for each of them, and find the delta with the original optimal compression of m. This gives us a ΔH_k for each continuer m_k. The probability for each of these, assuming binary bits, is

$$\begin{aligned} p(m_A|m) &= \frac{1}{Z} 2^{-\Delta H_A} \\ Z &= \sum_k p(k) \end{aligned} \tag{8.37}$$

Hence, there are determinate, objective algorithmic probabilities for ASU-DFT branches that can be limit-computed from the fixed-precision amplitudes of the preferred-basis superposition. The probabilities are based on bit-lengths, but can be computed from a fixed-precision representation without counting the bits of the amplitudes in that representation:

$$\begin{aligned} \Delta H_A &= L\left(\mathcal{F}_{min}(m_A)\right) \\ p(m_A|m) &= \frac{1}{Z} 2^{-L(\mathcal{F}_{min}(m_A))} \end{aligned} \tag{8.38}$$

where

$$\mathcal{F}_{min}(\mu) \quad = \quad \mathcal{F}(\mu, R_{min}) : \; R_{min} = \min\left(R = 1, 2, \ldots N : \mathcal{F}^{-1}(\mathcal{F}(\mu, R), N) \Leftrightarrow \mu\right) \qquad (8.39)$$

and $m_1 \Leftrightarrow m_2$ means that m_1 and m_2 describe equivalent conscious mental states.

So $\mathcal{F}_{min}(m_A)$ is the resulting amplitude list for the minimum resolution DFT on m_A that successfully regenerates m_A on decompression. This can be calculated by repeatedly performing higher and higher resolution DFT compressions on m_A (for $R = 1, 2, \ldots$) until one is reached that successfully regenerates the conscious state m_A.

Thus, there exists a function f such that

$$p(m_A|m) = \frac{1}{Z} f(m_A) \qquad (8.40)$$

where Z is the normalization constant.

This is state supervenience, from which noncontextuality follows. By additivity, all the other outcomes can be grouped into $m_{\neg A}$, the complement of m_A, for purposes of probability calculations, and it therefore does not matter what basis we consider m_A to be part of, and noncontextuality is proved. □

By assuming strong DFT, we can restrict ourselves to strict DFTs, and ignore the possibility that there are other more optimal compression methods. There is thus no need here to compare anything with m, or with m_B, just to get the prenormalized measure. Therefore, $f()$ really does depend only on m_A and not on the state it branched off from, or the alternative outcomes. Hence a conscious state's (prenormalized) measure depends only on *that* state.

Contextuality is not only ruled out in ASU-DFT, it does not even really arise as a possibility, except in a very derivative sense. The very idea of contextuality is an artifact of other conceptions of the wavefunction, since it effectively means that the probability rule can be a different rule, given the exact *same* situation (analytically), merely because of the way the situation is conceptualized as a "measurement". It seems hard to imagine how this could be the case in *any* ASU-based system, even without the strong DFT assumption. Hence, although the above proof does not follow through as straightforwardly under the weak DFT hypothesis—since we don't actually have the relevant algorithm at hand—it would be difficult to argue that it would not.

To create a version of the proof for the weak-DFT hypothesis (or for any other lossy compression algorithm, for that matter), we simply replace $\mathcal{F}()$ and $\mathcal{F}^{-1}()$ with analogous functions for whatever compression and decompression algorithms we are using. For the above proof *not* to carry through would require the compression algorithm itself to be "contextual". In other words, the result of compression and/or decompression would have to be different depending on one's choice of POVM.

But, as we have noted, the whole idea that an observer has a choice of basis or POVM runs counter to the ASU framework. Under ASU, no overt "measurement" needs to even be made, nor even any overt "observation". So the proof of noncontextuality is primarily a result of the fundamental assumptions ASU makes, not the properties of the strong DFT.

We can even proceed here without the assumption that there is any environment at all to "observe" or "measure", so the whole idea of observation or measurement is entirely optional (although we will need it to discuss decoherence and thermodynamic stability, but these are not required to justify the Born rule). Even if we simply have alternate mental states (in a solipsistic universe) that are outcomes of m, without an environment, there will still be nontrivial probabilities from the perspective of m. This is because there will still be a preferred POVM in which the effects correspond to distinct mental (macro)states. These states will have some $\triangle H$ bits when encoded on their own, which will determine their probabilities. This is the most essential form of any ASU probability analysis. We can layer on top of this the idea that there are one or two (or more) simultaneous measurements going on, or "intended" by the observer, but this is not the source of probabilities, nor does it affect the objective probabilities. Remember that under our synthetic-objectivist account of probability, even a bunny rabbit in a box, in the corner of the lab, has these probabilities—just by being the bunny rabbit that he is, without any awareness of what is involved or going on. There is no possibility for contextuality in this situation. The context is objectively (and synthetically) determined by the fact of the unity of consciousness.

Theorem 8.7. ***Amplitude dependence***: *given that* $|\psi_m\rangle = \sum_{k=1}^{n} c_k |k\rangle$ *is a symmetry transform of an optimal DFT compression of a conscious mental state* m*, in an ASU preferred basis* $\{|k\rangle\}$ *of resolution* n*, there exists a function* $f()$ *that returns the measure or prenormalized probability of an outcome, given only its amplitude:*

$$p(m_A|m) = \frac{1}{Z} f(c_A) \tag{8.41}$$

where Z *is the normalization constant and* A *is the index in the preferred basis for outcome* m_A*, and* c_A *is thus the amplitude for outcome* m_A*.*

Proof. We know already that there exists a function f' such that

$$p(m_A|m) = \frac{1}{Z} f'(A, c_1, c_2, \ldots, c_n) \tag{8.42}$$

since all the requisite information is encoded in the amplitudes $\{c_k\}$, given the index $k = A$ to the outcome in question. With (8.38) this gives

$$\begin{aligned} f'(A, c_1, c_2, \ldots, c_n) &= 2^{-L(\mathcal{F}(m_A))} \\ f'(A, c_1, c_2, \ldots, c_n) &= f''(m_A) \end{aligned} \tag{8.43}$$

where

$$f''(m_A) = 2^{-L(\mathcal{F}_{min}(m_A))} \tag{8.44}$$

But since $f'()$ depends only on index A and the amplitudes $\{c_k\}$, it follows from noncontextuality (Thm. 8.6) that there exists a function $f()$ such that

$$f''(m_A) = f(c_A) \tag{8.45}$$

and so with (8.44),

$$f(c_A) = 2^{-L(\mathcal{F}(m_A))} \tag{8.46}$$

and with (8.38)

$$p(m_A|m) = \frac{1}{Z} f(c_A) \tag{8.47}$$

and amplitude dependence is proved. □

Theorem 8.8. ***Synthetic-Unitary Probability:*** *Given a generative view of probabilities based on algorithmic synthetic unity, and assuming the strong DFT hypothesis, the algorithmic probability of any term in a superposition describing an optimally compressed conscious state, or of any term in a DFT of the same, must be proportional to its norm-squared amplitude.* ***[Quantum Analogue: Postulate #5, Born rule]***

Proof. Version 1. Immediate from Gleason (Thm. 3.27) and noncontextuality (Thm. 8.6). □

Proof. Version 2. Immediate from Everett stage 1 (Thm. 3.6) and amplitude dependence (Thm. 8.7). □

What we have done here is probably best characterized as embedding Gleason's (formal) proof in the synthetic context of ASU, thereby providing a defence of Gleason's sole synthetic axiom (noncontextuality). I have phrased this as a theorem above, but it is not really a purely formal proof.[69] The whole context we are working in here cannot at this point be fully formalized, so we have really only presented an informal defence of ASU noncontextuality. This can perhaps be made more formal, but not completely formalized so long as consciousness itself is not completely formalizable.

The fact that we can derive the Born rule either via Gleason and noncontextuality *or* Everett stage 1 and amplitude dependence is not surprising. Recall that [135, Lm.3] already derives Gleason's result by first deriving amplitude dependence from noncontextuality, and that Everett proves the Born rule from amplitude dependence. Version 1 (the Gleason version) will probably give a more

[69] On the other hand, there is really no such thing as a "purely formal" proof, since merely by calling something a "proof", we automatically imply an external propositional semantics for it. The formal system, on its own terms, is just a computation or program. Nonetheless, some "proofs" have far greater synthetic gaps than others, and we can apply the "formal" label as a relative, not absolute, property of a proof.

formal overall result, since the move from noncontextuality to amplitude dependence is completely formal, whereas in Version 2, it is couched in ASU terms. On the other hand, Version 2 may be easier to follow, and certainly gives a more intuitive explanation of probability within the ASU framework, given that more of the proof is couched in ASU terms.

In summary, then, unless our ASU-DFT system is guilty of some outright inconsistency, then noncontextuality holds and Gleason's theorem proves that the only way to incorporate those extra $\triangle H$ bits into an m they have been compressed out of, is for them to show up as the norm-squared of the amplitude values (and not as localized bit-lengths).

8.2.5 Synthetic-Unitary Measurement [Quantum Analogue: Postulate #3]

ASU-DFT has no built-in notion of measurement (observation of external variables). Branching occurs at objectively determined times and the branch-count has an objective value after a branching. However, it is deceptive to call this process "measurement". "Observation" is closer—and I therefore tend to prefer that word—but even this can be deceptive, since the only thing that is required is for the optimal program to generate multiple future mental states. No explicit observation of something in an external environment is required.

However, given that we have assumed a complex environment ($R_m \ll N_m$) in order to achieve some level of cosmic stability, it is natural to suppose (but I will not prove) that multiple future selves will have differing variables in an effective "environment", and that we should be able to model the process of explicit conscious observation of external variables, which I will call "measurement".

Recall that a branching occurs due to a unitary transform

$$|m\rangle \Longrightarrow \hat{O}\,|m\rangle = |m'\rangle = \sum_i c_i\,|m_i\rangle \tag{8.48}$$

that increases the local dimensionality from N_{t_m} to $N_{t_{m'}}$:

$$N_{t_m} < N_{t_{m'}} \tag{8.49}$$

where each $|m_i\rangle$ is a distinct mental continuer state of $|m\rangle$.

Definition 8.9. By "environment-free" observation, we will mean the case where

$$N_{t_m} = L(b) \tag{8.50}$$

where $L(b)$ is the localized measure of the number of bits in the brain $|b\rangle$ of $|m\rangle$, so that the compression algorithm compresses $|b\rangle$ into $|m\rangle$.

Clearly, if the compression algorithm that takes us from $|b\rangle$ to $|m\rangle$ were lossless, we could only have environment-free observation. But since the algorithm is lossy, we actually have

$$N_{t_m} \geq L(b) \tag{8.51}$$

Definition 8.10. By "environmental observation", or "measurement" we will mean the case where

$$N_{t_m} > L(b) \tag{8.52}$$

where any bits that we can throw away, while retaining a full description of m, are local "environment" or "environmental" bits.

If we were only concerned with decompression, we could always throw out the extra environmental bits, and consider them an artifact of our decompression algorithm (mere temporary storage or "working memory"). However, in order to maintain symmetry, our decompression needs to be reversible, and must retain these bits.

There is no way to rule out the possibility, from mere experience, of environment-free observation—essentially the solipsistic option ("my mind is all there is to the universe"). This would not even violate the ASU cosmic stability requirement. It is possible that objects of our conscious experience might all be required aspects of a local description of our brains, while still behaving in a stable and orderly fashion, without the necessity of considering them as part of an "external" environment.

However, given that ASU clearly allows for an environment, and given that the solipsistic option is not really practical, we will assume

Assumption 8.11. *The Environmental Assumption (Denial of Solipsism): measurement, or environmental observation* ($N_{t_m} > L(b)$)*, is more probable than environment-free observation* ($N_{t_m} = L(b)$).

Given this assumption, we can factor the Hilbert space $\mathcal{H}$ into $\mathcal{H}_b \otimes \mathcal{H}_v \otimes \mathcal{H}_e$ so that we have:

$$|\psi_m\rangle \quad = \quad |b \otimes v \otimes e\rangle \tag{8.53}$$

undergoing a unitary transform $\hat{O}$ that generates the measurement:

$$|m\rangle = |b \otimes v \otimes e\rangle \Longrightarrow \hat{O}\,|m\rangle = |m'\rangle = \sum_i c_i\,|m_i\rangle = \sum_i c_i\,|b_i \otimes v_i \otimes e_i\rangle \rightarrowtail |b_A \otimes v_A \otimes e_A\rangle \tag{8.54}$$

where $|b\rangle$ and the $\{|b_i\rangle\}$ are localized brain-states in $\mathcal{H}_b$, $|v\rangle$ and the $\{|v_i\rangle\}$ are environmental observables in $\mathcal{H}_v$, and $|e\rangle$ and the $\{|e_i\rangle\}$ are environmental states in $\mathcal{H}_e$. The original mental state is $|m\rangle$ and the $\{|m_i\rangle\}$ are the branched mental states. The pre-compression representation of $|m\rangle$ is $|b\rangle$.

We can distinguish between $|v\rangle$ and $|e\rangle$, for the time being, in whatever way is convenient. However, presumably if we had a correct model of consciousness, it would be evident how to do this. Hence, applying the assumption of servomechanism equivalence, I will tentatively call $|v\rangle$ the "controlled variable", presuming it to be the controlled environmental variable of a negative feedback control system or servomechanism [9, 161, 231]. Since any organism is controlling for any number of environmental variables at once, what we choose as the controlled variable can be a matter simply of what we are interested in describing ("we" meaning theorists modelling the conscious observer, not the conscious observer herself).

In environmental branching, it will always be possible to divide the environmental bits between some $|v\rangle$ and the remainder $|e\rangle$. Since there *are* extra environment bits, and we know they are required to maintain symmetry, it follows that any future evolution of the disturbances to the controlled variable $|v\rangle$ will depend on $|e\rangle$. If no such separation between $|v\rangle$ and $|e\rangle$ were possible, then it would follow that the entire external environment of the observer was under the observer's control. Yet, if this were so, there would be no justification for not including all those extra bits in the local description of the brain, in the first place (in other words, as part of $|b\rangle$). Hence, if there is any external environment at all (*i.e.* if we are to deny solipsism), then we must be able to decompose the environment into those bits under control[70] $|v\rangle$ and those that are not $|e\rangle$.

Probabilities for measurement outcomes are, under this model, straightforwardly derivable from the objective branch probabilities. The available objective branches are:

$$W = \{|m_i\rangle = |b_i \otimes v_i \otimes e_i\rangle\} \tag{8.55}$$

where, for any two distinct branches $|m_i\rangle$ and $|m_j\rangle$, $i \neq j$:

$$|b_i\rangle \neq |b_j\rangle \tag{8.56}$$

since otherwise, we will have violated the definition of an ASU macrostate. However, it is perfectly possible that we could have $|v_i\rangle = |v_j\rangle$ or $|e_i\rangle = |e_j\rangle$.

Definition 8.12. A "measurement partitioning" or "observation partitioning" of $\{|m_i\rangle = |b_i \otimes v_i \otimes e_i\rangle\}$ is defined as

$$\{|o_k\rangle = |b^o_k \otimes v^o_k \otimes e^o_k\rangle\} \tag{8.57}$$

[70] Note that "under control" does not mean completely *successful* control at all times. If I am driving my car, and trying to keep the centre of the road at the centre of my field of vision, I am probably going to be constantly correcting for disturbances that push my perception of the road's centre away from the centre of my vision. It does not follow that I am going to be 100% successful at performing such control. However, my perception of the road's centre remains the controlled variable, whether or not I have managed to achieve such control. Hence, those environment bits that are required for describing this variable will be part of $|v\rangle$ and not $|e\rangle$ (they will not float back and forth between the two, as I go in and out of successful control).

where for all pairs of distinct k values, k_1 and k_2 ($k_1 \neq k_2$), the brain states and controlled variables are distinct:

$$\left|b^o_{k_1}\right\rangle \neq \left|b^o_{k_2}\right\rangle \tag{8.58}$$

$$\left|v^o_{k_1}\right\rangle \neq \left|v^o_{k_2}\right\rangle \tag{8.59}$$

and derived from the original macrostate branches $W = \{|m_i\rangle\}$:

$$|v^o_k\rangle = \sum_{|m_i\rangle \in V_k \subseteq W} c_i \, |v_i\rangle \tag{8.60}$$

$$|b^o_k\rangle = \sum_i c_i \, |b_i\rangle \tag{8.61}$$

Principle 8.13. ***Principle of Synthetic-Unitary Measurement:*** *Denying solipsism, we model a measurement (environmental observation) by a measurement partitioning $\{|o_k\rangle = |b^o_k \otimes v^o_k \otimes e^o_k\rangle\}$ of an environmental observation $|m\rangle = |b \otimes v \otimes e\rangle \Longrightarrow \sum c_i \, |m_i\rangle = \sum c_i \, |b_i \otimes v_i \otimes e_i\rangle$. This is simply a coarser decomposition of our Hilbert space, based on the practical constraints of our experimental setup (we cannot actually distinguish the true macrostates in practice, anyway). The probabilities are directly derived from the objective branch probabilities:*

$$p(o_k|m) = \sum_{|m_i\rangle \in V_k} p(m_i|m) \tag{8.62}$$

and are themselves hence also objective. The measurement thereby entails a subjective "measurement collapse" (based on the objective "branching collapse"):

$$|m\rangle \Longrightarrow \hat{H} \, |m\rangle = |m'\rangle = \sum_k c^o_k \, |o_k\rangle \rightarrowtail |o_A\rangle \tag{8.63}$$

The unitary measurement transformation is a global (cosmic) transformation from one mental state to its continuers. ***[Quantum Analogue: Postulate #3, Observability]***

The observer is interested in a particular controlled variable, so she will be interested in modelling $\hat{H}$ in terms manageable to her, as a simpler operator, call it $\hat{H}_v$, describing the local dynamics of the controlled variable under a certain experimental setup. The result of a measurement should be real-valued (since we presumably started with $|b\rangle$, prior to compression, as a real-valued data structure), so we are permitted to require $\hat{H}_v$ to be Hermitian, with its eigenvalues representing the real-valued possible experimental outcomes. Of course, a larger analysis is possible, in principle, that will also model the full environment $\hat{H}_e$ as well as the observer $\hat{H}_m$, as well as all the interactions between these three, $\hat{H}_{\text{int}}$:

$$\hat{H} = \hat{H}_v + \hat{H}_e + \hat{H}_m + \hat{H}_{\text{int}} \tag{8.64}$$

Again, it is not strictly necessary to introduce such a measurement principle into ASU at all. It functions perfectly well without one. There is no inherent "measurement process", and the branching principle is sufficient to characterize transformations that induce local entropy increase from the

perspective of a given conscious mental state, and to derive objective probabilities. However, the above model of measurement seems reasonable (assuming non-solipsism) and is consistent with one widely used in quantum mechanics, while remaining compatible with ASU-DFT. It should be possible to interpret $\hat{H}_v$ as $\hat{H}_{\text{obl}}$, $\hat{H}_m$ as $\hat{H}_{\text{obr}}$, and $\hat{H}_e$ as $\hat{H}_{\text{env}}$ under the usual quantum mechanical model of measurement, possibly including, if desirable, environmentally-induced decoherence to characterize the possible stable outcomes of the measurement.

8.3 Discussion

While the previous section does not give us a full proof of the Born rule from the MWI assumptions alone, I think it does derive the Born rule within a certain context, simply by showing that

1. ASU-DFT provides a coherent and elegant overall interpretation of quantum mechanics, with the ability to derive most of the structure of the theory from *a priori* rationalist philosophical principles, along with some reasonable educated guesses consistent with those principles; and, more specifically, that
2. Under this framework, Gleason contextuality does not arise as a possible interpretation of measurement, or of quantum probability, and hence the Born rule follows, meaning that
3. We have completely side-stepped the issue of convergence in the infinite limit—and this is, in fact, completely acceptable within the ASU-DFT framework, which is neither frequentist nor supportive of branch-counting.

While a strong case can be made that something like Gleason noncontextuality follows from ASU *without* the DFT hypothesis, it is difficult to frame such a result as relevant to the Born rule debate, since the analytic context of such noncontextuality would be broader than the analytic context of quantum mechanics. Using the toy examples from Ch. 6, without some connection to the analytic quantum postulates (such as the strong-DFT hypothesis provides), would only serve as an analogy for the quantum case. The purpose of the toy examples is to frame some of the general features of quantum probability in the context of a more general theory of probability for observers in formal systems. However, an actual derivation of the Born rule clearly requires that this general framework be connected to the Hilbert space context of quantum mechanics.

But keep in mind that, while I earlier described the strong DFT hypothesis as under-justified, that was in the context of an *a priori* derivation of *all* the quantum postulates. The strong hypothesis is perfectly acceptable in the context of the MWI Born rule debate, given that we already know empirically that the wave equation *is* describable as a Fourier transform. Nonetheless, I have tried to derive as much of the first four quantum postulates as I can *a priori* from ASU principles—at least in outline—in order to make the broader ASU framework as strong and compelling as possible.

Nonetheless, for purposes of the Born rule debate, we do not need to go this far—we are, after all, allowed to assume the first four postulates in responding to the Born rule objection.

It might also be possible to cite Wallace's proof instead of Gleason's or Everett's, since my proof of noncontextuality includes a proof of state supervenience (8.40). We would then merely need to dismiss branch-counting within the algorithmic framework to prove the Born rule (and I have already, I think, effectively done this). In addition, perhaps other of Wallace's axioms could be strengthened or eliminated by appealing to ASU (although I have already argued in §3.3.10 that the other axioms are largely benign). Essentially, we would be using ASU-DFT here as a way of supporting weak points in Wallace's justifications for his axioms, and I believe an argument based on Wallace's proof is thus not much more than an exercise for the reader, and would look pretty similar to the argument based on Gleason's proof. Nonetheless, Gleason's proof is cleaner, with fewer synthetic assumptions and no reliance on controversial decision-theoretic formulations of probability (I am already assuming my own controversial assumptions about probability here, so best not to muddy the waters by introducing still more!).

Of course, none of this negates the possibility of a refutation of ASU-DFT based on the derivation of some *other* non-Born *and* non-branch-counting measure. Such a demonstration would revive the Born rule objection in an algorithmic context, as it would mean that the framework itself is inconsistent with the Born rule. And, given Gleason's proof, and the fact that noncontextuality is a natural consequence of ASU-DFT, this would imply that ASU-DFT itself was inconsistent. A possible path to such an inconsistency might be to prove that the optimal compression of a conscious state cannot be a DFT. However, the onus here is on the objectors to show an inconsistency between Gleason and noncontextuality, on the one hand, and ASU-DFT on the other.

Interestingly, there seems not much room here for the frequentist-based proofs, such as Everett stage 2 and its relatives. Under the ASU-DFT framework, there is simply no reason to take this approach. We are not being frequentists here, so there is no reason we need to be compelled by our principles to invoke infinite limits of repeated observations, in order to talk about probabilities. Neither do we reject the idea that a single case can have objective probability (because of the existence of open cases), so there is no reason not to simply use a classical counting method, which is essentially what we have done. Hence, it is sufficient to show that the norm-squared amplitude measure is the only possible analytic measure on the space, given that the issue of noncontextuality does not arise, for synthetic *a priori* reasons. We have no reason to take branch-counting seriously in this framework, so there is simply no motivation to invoke infinite sequences, merely to force norm-squared measures to equate with branch-counting.

I am not claiming everyone need accept the ASU foundation, of course. Those who don't will still perhaps have a viable Born rule objection at their disposal, or a reason to support the frequentist approach. However, they will need, I think, to be able to point to those parts of the ASU-DFT framework that they object to, and why they still think that there is some other valid synthetic *a priori* way of approaching objective probability. Simply assuming without justification, as has been done in the past, that some form of branch-counting is the appropriate synthetic *a priori*, is no longer sufficient—I believe I have defended the algorithmic alternative at enough length to compete more than favourably with branch-counting. I would hope that any future attempts to invoke the Born rule objection will address these issues, and I would challenge any would-be Born rule objectors to meet my challenge to do at least one of the following:

1. Present a foundation for synthetic *a priori* calculation of probabilities for observers emergent from an analytic (formal) system that is at least as defensible as the algorithmic foundation I have presented. Branch-counting is permitted here, but it must be defended with rational argument: *why* should simple observer or world counts correspond to probabilities? Why should branches matter? A reasoned argument for the Thirder solution to Sleeping Beauty might be a helpful starting point.

2. Show why the DFT assumption is invalid, given the rest of the ASU framework, or (in the spirit of #1) show that the general ASU framework leads more naturally to some *other* compression scheme than the DFT (the onus will then be on me to see if this new scheme is consistent with the Born rule).

3. Show that, even given the DFT assumption, the ASU framework does not, in fact, lead to the Born rule. This could take the form of showing an actual error in my reasoning, but might also consist of simply showing that one of the numerous gaps in my derivation is untenable. I recognize that my "derivation" is not entirely formal. It has numerous gaps that I feel I have argued persuasively for, at least from within the ASU perspective, but there certainly remains plenty of room here to find problems! A would-be objector may be able to show that one of these gaps is fatal or in some way even contradictory with some aspect of the ASU framework. Perhaps a *reductio* could even be performed that would show that the framework as currently conceived is inconsistent, by proving some other rule within the same system, with at least as much persuasiveness.

4. Argue against the ASU framework itself, without contradicting Everett's three primary philosophical assumptions. I have argued that the ASU framework can be considered a natural outgrowth of Everett's assumptions, but there is certainly potentially room here for an objector to argue that I have made unjustified leaps in my reasoning and that Everett's assumptions are equally (or more) compatible with some other competing way of conceiving conscious observers that emerge from purely analytic systems. While answering this kind of objection goes beyond my goals in this dissertation—which mostly just assumes the general ASU foundation—such objections are relevant and can be addressed within the context of Everett's interpretation.

5. Argue against Everett's three primary philosophical assumptions (wavefunction realism, psychophysical parallelism and servomechanism equivalence). While I list this possible avenue of objection for the sake of completeness, I do not consider such objections all that relevant to the concerns of this dissertation. I think it is pretty clear that if one is willing to reject Everett's primary three assumptions, that algorithmic synthetic unity will probably go down

in flames quite readily. It was never my intent to argue against such positions, which are based on fundamentally different metaphysical and epistemological viewpoints—about basic things like physicality, consciousness, probability and analysis—in a way that is clearly opposed to my own philosophical starting point. Such arguments go well beyond differences of opinion on quantum foundations, however, and objections in this category are best dealt with by retreating from quantum mechanics into the debate over these more fundamental metaphysical and epistemological differences, which need to be settled before useful debate on quantum foundations or the Born rule objection can proceed.

9 Conclusions

9.1 Summary

Everettian Born-rule proofs, mostly frequentist in nature, are all subject to some combination of objections based on maverick worlds, hidden assumptions of amplitude dependence, or problems with infinities.

I have argued that the issues surrounding maverick worlds are tractable, both when those worlds are a quantitative concern ("do they mess up our probabilities?") and when they are a moral or aesthetic concern ("do they mess with our minds?"). For the former concerns, we found that the problem was mostly an over-reliance on the *a priori* validity of world-counting, which we found was poorly justified. Once we rejected this, there was no obvious problem with the very low-amplitude maverick worlds. But, for those who still feel uncomfortable with the very idea of crazy maverick worlds really existing out there somewhere, we saw that it really doesn't even make sense to call these possibilities "worlds" in the first place, as it makes no sense to even imagine someone inhabiting one of them.

The problems with infinity are, I have argued, very real—and the frequentist proofs all fail due to this issue. When such a proof works only in the infinite limit, or only for infinity itself, but fails (even approximately) to work in the case of very large but finite n, then there is something fishy about the idea that anything substantive has been proved at all.

Problems with amplitude dependence, resulting in circular reasoning for many proofs, are the most difficult to get rid of. Even the problems with infinity can be viewed as amplitude dependence in disguise, since they are providing an excuse for tossing low amplitude worlds out. Wallace's proof, which does not seem on the face of it to assume amplitude dependence, arguably does so in a weak fashion, by assuming that branch-counting—the main competitor to amplitude-counting—is untenable, and by assuming state supervenience, which we have seen is very closely related to amplitude dependence.

Branch-counting is at the crux of all these attempted proofs. Eliminate the assumption that counting branches makes sense, and you eliminate the need for the frequentist proofs, and you also

eliminate whatever weak justifications may have existed for disregarding Gleason's proof. Hence, I have argued that the best approach to a Born rule proof is to show that one's overall interpretation of quantum theory is consistent and implies noncontextuality and/or amplitude dependence. I believe I have made a good case for these points, within the ASU-DFT framework.

There will, of course, be those who remain fixated on the idea of a "formal" Born rule proof, with no assumptions. It seems that MWI Born rule proofs keep getting more complicated to answer objections, but will never end up entirely answering them, because there is no absolute way to derive probabilities for perceptual observations without *some* kind of assumptions about how the perceiver's mind interfaces with the system (unless you have a complete and uncontroversial model of what constitutes an "observation", and likewise, "perception", "consciousness" and "mind").

I believe I have shown, in outline, how these fundamental postulates emerge naturally out of ASU, so long as we make the one additional assumption that the optimal compression of a conscious state is a discrete Fourier transform. This one (I think, very well-motivated) extra assumption provides a means of introducing the Hilbert space representation, and its model of observation and branching/collapse. This does not *prove* that the wavefunction ontology is algorithmic, but it is striking evidence. I do not know of any other interpretation of quantum theory that can derive anywhere near this much of the theory from *a priori* philosophical principles, even with the kind of gaps and assumptions I have allowed. And this *a priori* convergence with the empirical theory lends credibility to my claim that Gleason's proof provides ASU-DFT with the appropriate justification for the Born rule.

Some may still object, as is often claimed, that Gleason's proof lacks "physical insight" into the emergence of probability. However, my scepticism over the coherence of the conventional notion of the "physical" should, by this time, be clear to the reader. I prefer to view the physical as something that is emergent from computation, simply because computation is something that can be formally defined, and appears to be maximally expressive. So "physical relevance" is something that will have to be defined and clarified before I will consider that Gleason's proof needs to deliver more of it. In the meantime, I have no need for it.

But an objector might also claim that Gleason's proof lacks relevance for the ASU, as well, since it does not seem to be about counting programs. However, recall that, since we expect programs to have large mutual information, and because we are explicitly deriving probabilities from a compressed representation, we do not expect to be counting out bits in this representation. We expect, in fact, not to be doing so, for the reasons explained earlier: probabilities derive ultimately from bit counts of optimal compressions that are not locally available to us. (Even if we could imagine accessing this optimal representation directly, our probabilities would be based on the delta entropies defined

in the last chapter, between *two* such representations, not the bit count of a single compressed representation.) If we are to access such entropies by examining the wavefunction as it appears to us, we won't be counting bits in the representation of the alternative outcomes. There is thus nothing inconsistent, or even "lacking in relevance" about presenting a proof that there is only one measure consistent with ASU-DFT *and* with our ΔH entropy count. Since it follows from Gleason that only amplitude-counting is consistent with both of these things, this should be sufficient, so long as we are committing no inconsistency.

9.2 Future Work

9.2.1 Detailed Examples

While I have argued for the adequacy of Gleason, assuming consistency of ASU-DFT, there is no doubt that a fully worked-out example would be more convincing. It is possible that there *is* some inconsistency in the ASU-DFT framework, in which case my argument for Gleason's proof does not amount to much. This is why an actual fully worked out example is called for. If a simulation, in the form of an actual program, could demonstrate the ASU-DFT system, and compute and demonstrate norm-squared probabilities, we would have both a consistency proof (in the form of a program) and the analytic *a posteriori* evidence of being able to actually run the program and compute the resulting probabilities. This would leave little room for doubt, as it would be a constructive proof, in terms of program counting directly. Presumably, such a demonstration could readily be generalized on paper to an *a priori* something like our proof in the last chapter—and the process of doing this would probably aid in making the *a priori* version much tighter and more formal.

The model of consciousness used in such a simulation need not be much more detailed than what we have already used. So long as the optimal compression of our model *was* actually a DFT—or even plausibly a DFT—then that should be adequate. Recall that in general one cannot prove that one has achieved the optimal compression, anyway, so strict optimality can only be expected for very restricted examples.

9.2.2 More Accurate Models of Consciousness

Another possible phase for this research is to take the detailed simulation mentioned in the previous section and actually try to incorporate a serious model of consciousness. For the very first pass, the "conscious" observer model might simply be a single bit that is modified periodically by a simple disturbance, and then brought back to a steady state by a simple servomechanism. While hardly an adequate model of consciousness, it does reflect one of the most basic characteristics of living

things: negative feedback, closed loop control [161, 9, 231, 140, 162, 163]. More complex models of such systems can be introduced, as we feel the need for more complexity and more realism.

It is completely unknown, given the current state of our understanding of ASU, how much detail about consciousness will be helpful in deriving an *a priori* physics. It is possible that even a simple feedback model is too much detail. Or perhaps much more detail will be needed. It is also possible that the entire question is intractable. More research is needed.

It is also quite possible that very few details are necessary to address the interpretation of nonrelativistic quantum mechanics—and that simply assuming DFT-compressibility is the best we can do—while at the same time, more details might be very helpful in extending ASU past the five postulates and into the actual details of general relativity and particle physics. There is no reason not to look for *a priori* synthetic derivations in these areas, as well. Either way, attempts to build such models, and model ASU within them, will help us figure all this out.

I will not develop these ideas here in any further detail, but will simply list some basic features needed for such computer models—even in their initial preliminary implementation—in order for them to be meaningful to the issues we have been addressing:

1. The model must include, or at least allow for, both an environment and an observer.
2. The observer must have some state that reaches a value dependent on changes in the environment.
3. A language that allows bit-counting needs to be given for encoding observer and environmental states into potentially shorter descriptions using Fourier-based compression.
4. The universe (meaning observer+environment) must have a description in this encoding language that is at least at times shorter than the description of the observer state alone (this will allow for an effective stable "environment", and so is related to requirement #1).
5. The model should allow for actual computer simulation in order to test ideas and further explore its properties.

Note that I have not included "superpositions," or "interference effects", or "unitary evolution", since the hope is that the model will illustrate how these features develop naturally and inevitably out of any system with the above properties, in the same way I have tried to sketch out in Ch. 7-8. Ideally, we would like to not even assume the DFT hypothesis. However, given that optimality of compression is uncomputable, an assumption about optimality may perhaps always be necessary, at some point. Nonetheless, we need to strive to make these assumptions as weak as possible, and it is possible that we can go much further than I have in this dissertation towards justifying the strong DFT hypothesis (or at least something similar that will deliver the analytic postulates at least as readily). While justifying it for real human minds is still too large a problem, it may be possible to

build a good case for the hypothesis for some small, but well-motivated, model of a control system displaying life-like properties.

9.3 Concluding Remarks

Finally, I would like to present a challenge to any group of two or more individuals who wish to engage in a debate over the Born rule objection (and perhaps quantum probability in general). First, agree to put quantum mechanics aside for the initial phase of your debate. Examine instead the Sleeping Beauty probability puzzle, which has the same epistemic structure as Everettian quantum probability. Come to an agreement first on whether the Thirder or Halfer solution is the correct one. If you can decide on this, you have defined what you mean by "probability" sufficiently to continue on to a discussion of quantum mechanics. If you reach an impasse, then decide to have a two-pronged debate with both Sleeping Beauty solutions accepted in turn, as provisional axioms. Debate the Born rule objection, in other words, agreeing first that everyone shall be Halfers, and then carry on a second phase where everyone agrees to be Thirders. You may be surprised how much easier it is to come to agreement under each of these provisional axioms. Even if you do not end up agreeing in the end, you may still reach a rather satisfying understanding of each other.

Your debate may not in any way follow the lines of this dissertation, as it is possible none of the participants will share my computationalist leanings, and even if they do, their take on it may lead them in a different direction. However, my guess is that the Halfers will nonetheless look for the underlying generating mechanisms behind the phenomena, analogous to Lewis's coin flips, and leading to counts based on something that can be derived from the analytic structure of the wavefunction, such as amplitudes. Thirders, on the other hand, will consider something like Elga's subjective or synthetic indistinguishability to be the requirement for a countable, and will be apt to count something purely synthetic, such as branches, worlds or observers. There still may be disagreement, but the debate will be much more fruitful than if the participants attempt to delve immediately into quantum mechanics, with no idea whether they even share a common notion of what probability is.

The participants could decide to bypass the Sleeping Beauty puzzle, and go straight to an all-out debate over the general interpretation of probability, but this is such a huge and sprawling question, with so many nuances and different possible positions, that you may never return to quantum mechanics. The Sleeping Beauty puzzle has the advantage of encapsulating in a single, easy-to-understand gedanken experiment just the features of the general probability question that are most relevant to quantum probability—and it can even be set in a classical universe, if you like.

Of course not all Halfers will necessarily support algorithmic synthetic unity. However, Halfers

share certain common ideas about probability, and someone who supports algorithmic synthetic unity is, I believe, almost certain to be a Halfer. The ASU approach to Halferism takes the epistemic probability structure of the Sleeping Beauty probability problem and applies it straight-out to the ontic structure of the Everettian probability problem, to the extent that it assumes that the strange mixed-up "sequence" of events that Sleeping Beauty experiences, due to the amnesia, is (in the ontic case) the *actual* sequence of events. This is so because ASU takes the arrow of time and causal sequence of events to be fundamentally synthetic *a priori* in nature, rather than the result of physical or material "walls" that exist between subsystems or times. Time is not subjective in this view, but it *is* synthetic. In other words, what is only subjectively mixed up about Sleeping Beauty's "time arrow" would *literally* be mixed up in exactly the same fashion, if the structure of the Sleeping Beauty puzzle were really *all* there was to Sleeping Beauty's universe.

This viewpoint is subject, of course, to all the traditional objections to idealism, such as Johnsonian rock-kicking, even if we do not commit ourselves to all-out idealism. I indicated earlier that the algorithmic interpretation does not necessarily assume an idealist metaphysics, but does tend to suggest one—in particular something like transcendental analytic idealism. Hopefully, the reader can see more clearly now why this is so. It is difficult to motivate the idea of using Solomonoff probability to (essentially) randomly pick from all possible continuer programs, without considering that perhaps this means we are actually dealing with an ontology of all possible programs. We can sidestep the issue the way Everett sidesteps Strong AI, by using the "as if" qualifier: assume that quantum theory functions *as if* all possible programs exist. But then why would we not go the extra step of actually adopting some kind of computational idealism? There is surely little left for a physical materialism to do for us. Especially since ASU provides a plausible explanation, in idealist terms, for the stability of the universe—the kick-back of rocks—and the absence of maverick worlds.

It is, however, still possible that quantum mechanics may act "idealistically", while ultimately the universe more generally does not. I have argued that the postulates of quantum mechanics are *a priori*. Does this mean that, for instance, the charge of an electron is also *a priori*? That is a bigger leap to make, at this point, and while I believe there is real, direct evidence for the *a priori* nature of the quantum postulates, nothing like that exists for the fundamental constants of particle physics, or for general relativity. However, there *are* independent cosmological arguments for multiple worlds, independent of Everett, which may (or may not) point to an ultimate convergence of these ideas into a coherent and comprehensive *a priori* physical theory (as proposed, for example, by Max Tegmark [212]).

In addition, so much of what we know about physics is founded on quantum theory, that it seems that if such a theory appears itself to operate idealistically, it would seem, at least, that an

idealist approach to the rest of physics should not be considered the taboo subject that it has been for some time now. The more success we find for such methods, the more legitimacy it lends to the idea that a thorough-going metaphysical rationalism might, after all, be a viable overarching framework for the empirical sciences. I have advocated this rationalist approach to science elsewhere [168, 169, 172, 171, 174, 170], as have others [149, 212, 87, 191, 139].

But even for those who do not share my metaphysical leanings, I hope I have shown that the methods of algorithmic synthetic unity can greatly clarify some of the more vexing controversies in the foundations of physics, such as those surrounding the nature of collapse, amplitude interference and the Born rule. Perhaps this will spark more interest in systematic synthetic *a priori* approaches to physics, whether along the lines of ASU or something entirely different. Such methods fell into disrepute in the early twentieth century, before quantum theory even came on the scene[71]—and, in some ways, for good reason—as they came to be associated with the assumption that there are mysterious non-analytic essences at work in the universe (although it must be noted that even matter and energy are mysterious essences, if taken as metaphysical absolutes, leaving materialism in not much better of a position, ultimately). I hope I have demonstrated that one can take great advantage of the power of synthetic *a priori* methods, however, *without* assuming mysterious essences, irreducible syntheticity, or any underlying mentalistic conception of reality, and by neither dogmatically asserting, nor discounting the possibility, that there may ultimately exist a complete analysis—at least in principle—of everything.

But if even a tentative analytic idealism is to be adopted in physics, there needs to be something like the cosmic thermodynamic stability that the ASU provides (which is imposed by the highly compressible nature of consciousness) in order to answer Johnson's rock-kicking challenge. And if

[71] I will digress briefly to give my own opinion and perspective on the historical development (although I am not an expert in this area). As a matter of historical convenience, and to give the reader a bit of a grasp of the timelines involved, I will take the beginning of the age of the synthetic *a priori* as 1781, the year Kant published the *Critique of Pure Reason*. I will take the end of this era to be 1922 or 1925, with 1922 as the beginning of the Vienna Circle, with its explicit rejection of the synthetic *a priori*—and indeed all metaphysics—and 1925 as the year that F.H. Bradley died. Bradley [25, 26, 27] was the nineteenth century absolute idealist who championed an *a priori* many-worlds ontology; his work was considered by many at the time to have been the pinnacle of nineteenth century metaphysics, and he was a main target of the revolt against idealism lead by Bertrand Russell (originally an idealist himself), G.E. Moore and the Vienna Circle. Here is the historical irony. Metaphysical methods—especially those which recognized the synthetic *a priori*—were being rejected, largely on the grounds of their historical futility, and in favour of strictly empirical methods—at the very same time that *empirical* (synthetic *a posteriori*) methods were revealing that reality had a structure that was, in outline, essentially what the synthetic *a priori* methods of the absolute idealists had suggested all along.

But it was not until 1957 that anyone (Everett, to be precise) realized that this was the case, and by that time the synthetic *a priori* had long since been relegated to the historical trash-heap. The anthropic principle has resulted in a revival of interest in such methods, but without the mentalistic and quasi-mystical bent that prevented the absolute idealists from interfacing productively with empirical scientists.

The leap from the existence of synthetic *a priori* truths to the existence of unknowable essences never was justified, and remains unjustified today. Scientific advance requires the cooperative use of *a priori* and *a posteriori* methods. Those methods of science that recognize the synthetic *a priori*, such as the anthropic principle, quickly devolve into mysticism when the synthetic *a priori* becomes anything but a subset of the analytic, since it is analysis that yields all the consequences of the ultimate arbiter of truth as we know it: reason.

my brand of purist (one might even say extremist) Halferism is to fly, we must be able to view our conscious brains as mere local manifestations of a more fundamental global representation that both defines the observable universe (but *not* ultimate reality!) in terms of that consciousness, and insists that such a universe be fundamentally smaller and simpler than any local neurological description of the same brain could ever be.

The universe is, indeed, an undivided whole under this view, as David Bohm would have it [19]. But this "holographic" nonlocality is not some mystical feature of a wavefunction that guides, but is entirely unaffected by, our particulate bodies. Rather, the universe—meaning *physical* reality—is a transcendentally ideal, synthetic aspect of a larger, nonphysical and wholly *analytic* reality. Hence physical law, in this kind of idealism, will always be synthetic *a priori*, consisting of the simplest (shortest bit-length) descriptions of our experience, constrained by (and indeed existing within) an ultimate ontology that is computational in nature. And this ontology is computational—or at least must be taken as computational for the purposes of human enquiry—because it is computation that represents, so far as is known to-date, the widest possible spectrum of precisely thinkable thoughts.

Bibliography

[1] Wilhelm Ackermann. Zum Hilbertschen Aufbau der reellen Zahlen. *Mathematische Annalen*, 99:118–133, 1928. pages: 427, 438

[2] Wilhelm Ackermann. Zur widerspruchsfreiheit der Zahlentheorie. *Math. Ann.*, 117:162–194, 1940. pages: 440

[3] S Aerts. Quantum and classical probability as Bayes-optimal observation. *arXiv:quant-ph*, 0601138, 2006. pages: 19

[4] David Z Albert. *Quantum Mechanics and Experience*. Harvard University Press, Cambridge, 1994. pages: 13, 60

[5] David Z Albert. *Time and Chance*. Harvard University Press, Cambridge, 2003. pages: 142

[6] David Z Albert. Probability in the Everett picture. In *Many Worlds? Everett, quantum theory, and reality*. Oxford University Press, Oxford, 2010. pages: 23

[7] David Z Albert and B Loewer. Interpreting the many worlds interpretation. *Synthèse*, 77:195–213, 1988. pages: 13, 17, 125

[8] D M Appleby. Probabilities Are Single-Case or Nothing. *arXiv:quant-ph*, 0408058, 2004. pages: 225

[9] W R Ashby. *An Introduction to Cybernetics*. Chapman and Hall, New York, 1956. pages: 393, 402

[10] Siani Baker and Alan Smaill. A proof environment for arithmetic with the omega rule. In *Integrating Symbolic Mathematical Computation and Artificial Intelligence*, pages 115–130. Springer-Verlag, Berlin, Heidelberg, 1995. pages: 440

[11] Leslie E Ballentine. The statistical interpretation of quantum mechanics. *Reviews of Modern Physics*, 42(4):358–381, 1970. pages: 7, 128

[12] V Bargmann. Note on Wigner's Theorem on Symmetry Operations. *Journal of Mathematical Physics*, 5(7):862, 1964. pages: 82

[13] Chris Barker. *Iota and Jot: the simplest languages?* Http://semarch.linguistics.fas.nyu.edu/barker/Iota/, 2009. pages: 431

[14] H Barnum, Carlton M Caves, J Finkelstein, C A Fuchs, and R Schack. Quantum Probability from Decision Theory? *Proceedings of the Royal Society of London A*, 456:1175–1182, 2000. pages: 19

[15] Jeffrey A Barrett. *The Quantum Mechanics of Minds and Worlds*. Oxford University Press, Oxford, 1999. pages: 13, 125, 126, 332

[16] John S Bell. On the Einstein-Podolsky-Rosen paradox. *Physics*, 1(3):195–200, 1964. pages: 105

[17] John S Bell. *Speakable and Unspeakable in Quantum Mechanics*. Cambridge University Press, Cambridge, 1987. pages: 17

[18] John S Bell. The measurement theory of Everett and de Broglie's pilot wave. In *Speakable and Unspeakable in Quantum Mechanics*. Cambridge University Press, Cambridge, 1987. pages: 332

[19] David Bohm and Basil J Hiley. *The undivided universe: an ontological interpretation of quantum theory*. Routledge, London, 1993. pages: 105, 140, 406

[20] G Boole. *An Investigation of the Laws of Thought, on which are Founded the Mathematical Theories of Logic and Probabilities*. Walton and Maberley, London, 1854. pages: 222

[21] George Boolos, John P Burgess, and Richard C Jeffrey. *Computability and Logic*. Cambridge University Press, Cambridge, 5th edition, 2007. pages: 254, 258, 425, 434, 443

[22] Nick Bostrom. Self-Locating Belief in Big Worlds: Cosmology's Missing Link to Observation. *The Journal of Philosophy*, 99(12):607–623, 2002. pages: 304

[23] Nick Bostrom. Sleeping Beauty and Self-location: A Hybrid Model. *Synthèse*, 157(1):59–78, 2006. pages: 288

[24] James Boswell. *The Life of Samuel Johnson, LL.D.* John Sharpe, Piccadilly, London, 1830. pages: 332

[25] Francis Herbert Bradley. *Appearance and Reality: a metaphysical essay*. Swan Sonnenschein, London, 1897. pages: 405

[26] Francis Herbert Bradley. *Essays on Truth and Reality*. Clarendon Press, Oxford, 1914. pages: 405

[27] Francis Herbert Bradley. *Writings on Logic and Metaphysics*. Oxford University Press, Oxford, 1994. pages: 283, 405

[28] Luitzen Egbertus Jan Brouwer. Uber definitionsbereiche von Funktionen. *Math. Ann.*, 97:60–75, 1927. pages: 440

[29] Jeffrey Bub and Rob Clifton. A uniqueness theorem for 'no collapse' interpretations of quantum mechanics. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 27(2):181–219, 1996. pages: 102

[30] Jeffrey Bub, Rob Clifton, and Sheldon Goldstein. Revised Proof of the Uniqueness Theorem for 'No Collapse' Interpretations of Quantum Mechanics. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 31(1):95–98, 2000. pages: 102

[31] Jeffrey Bub and A Elby. Triorthogonal uniqueness theorem and its relevance to the interpretation of quantum mechanics. *Physical Review A*, 49(5):4213–4216, May 1994. pages: 102

[32] Roman V Buniy, Stephen DH Hsu, and AZee. Discreteness and the origin of probability in quantum mechanics. *arXiv:hep-th*, 0606062, 2006. pages: 169

[33] Mark S Burgin. Inductive Turing machines. *Notices of the Academy of Sciences of the USSR*, 270(6):1289–1293, 1983. pages: 439, 441, 446

[34] Mark S Burgin. Arithmetic hierarchy and inductive Turing machines. *Notices of the Academy of Sciences of the USSR*, 299:390–393, 1988. pages: 450

[35] Mark S Burgin. Universal Limit Turing Machines. *Doklady Mathematics*, 46(1):79–83, 1993. pages: 441

[36] Mark S Burgin. Super-recursive algorithms as a tool for high performance computing. In *Proceedings of the High Performance Computing Symposium Proceedings of the High Performance Computing Symposium*, pages 224–228. San Diego, 1999. pages: 439, 441

[37] Mark S Burgin and Yu M Borodyanskii. Infinite processes and super-recursive algorithms. *Soviet Mathematics - Doklady*, 44(3):800–803, 1992. pages: 441

[38] P Busch. Quantum states and generalized observables: a simple proof of Gleason's theorem. *Physics Review Letters*, 91(120403), 2003. pages: 199

[39] Peter Byrne. The Many Worlds of Hugh Everett. *Scientific American*, December 2007. pages: 16

[40] Peter Byrne. *The Many Worlds of Hugh Everett III: Multiple Universes, Mutual Assured Destruction, and the Meltdown of a Nuclear Family.* Oxford University Press, Oxford, 2013. pages: 16, 205, 333, 334, 345

[41] Brandon Carter. Large Number Coincidences and the Anthropic Principle in Cosmology. D. Reidel Publishing Co., 1974. pages: 306

[42] Brandon Carter. Anthropic interpretation of quantum theory. In *Interdisciplinary Colloquium La Philosophie de la Nature aujourd'hui?*, Paris, 2003. pages: 328

[43] Carlton M Caves. Subjective probability and quantum certainty. *Studies in History and Philosophy of Modern Physics*, 38:255–274, 2007. pages: 19

[44] Carlton M Caves, Christopher A Fuchs, Kiran Manne, and Joseph M Renes. Gleason-type derivations of the quantum probability rule for generalized measurements. *Foundations of Physics*, 34(2):193–209, 2004. pages: 197, 199

[45] Carlton M Caves, Christopher M Fuchs, and Ruediger Schack. Quantum probabilities as Bayesian probabilities. *arXiv:quant-ph*, 0106133v2, 2001. pages: 19

[46] Carlton M Caves and Rudiger Schack. Properties of the frequency operator do not imply the quantum probability postulate. *arXiv:quant-ph*, 0409144v3, 2005. pages: 19, 160

[47] Gregory J Chaitin. On the length of programs for computing finite binary sequences. *Journal of the ACM*, 13(4):547–569, 1966. pages: 243

[48] Gregory J Chaitin. On the length of programs for computing finite binary sequences: statistical considerations. *Journal of the ACM*, 16(1):145–159, 1969. pages: 243

[49] Gregory J Chaitin. A Theory of Program Size Formally Identical to Information Theory. *Journal of the ACM (JACM)*, 22(3):329–340, 1975. pages: 243

[50] Gregory J Chaitin. Algorithmic information theory. *IBM Journal of Research and Development*, 21:350–359– 496, 1977. pages: 243

[51] Gregory J Chaitin. Algorithmic Information Theory. Cambridge University Press, Cambridge, 1987. pages: 243, 453

[52] Gregory J Chaitin. *Lisp Interpreter Applet.* Http://www.cs.auckland.ac.nz/~chaitin/unknowable/lisp.html. Cited: 8 Nov. 2013, 2000. pages: 431

[53] Marvin Chester. *Primer of Quantum Mechanics.* Krieger Publishing Company, Malabar, 1992. pages: 60

[54] Alonzo Church. A set of postulates for the foundation of logic (part 1). *Annals of Mathematics*, 33:346–366, 1932. pages: 252, 430

[55] Alonzo Church. A set of postulates for the foundation of logic (part 2). *Annals of Mathematics*, 34:839–864, 1933. pages: 252, 430

[56] Alonzo Church. An unsolvable problem of elementary number theory. *American Journal of Mathematics*, 58:345–363, 1936. pages: 250, 252, 427, 430, 431

[57] Alonzo Church. *The Calculi of Lambda-conversion.* Princeton University Press, Princeton, 1941. pages: 252, 430

[58] F Crick and G Mitchison. The function of dream sleep. *Nature*, 304(5922):111–114, 1983. pages: 342

[59] Haskell B Curry. Grundlagen der kombinatorischen logik. *American Journal of Mathematics*, 52:509–536, 789–834, 1930. pages: 252, 255, 431

[60] Haskell B Curry. The combinatory foundations of mathematical logic. *Journal of Symbolic Logic*, 7:49–64, 1942. pages: 432

[61] Haskell B Curry and R Feys. *Combinatory Logic*, volume 1. North-Holland, Amsterdam, 1958. pages: 255, 431

[62] Haskell B Curry, J R Hindley, and J P Seldin. *Combinatory Logic.* North-Holland, Amsterdam, 1972. pages: 255, 431

[63] E B Edward Brian Davies. *Quantum theory of open systems.* Academic Press, London, 1976. pages: 201

[64] Pierre Simon Marquis de Laplace. *A Philosophical Essay on Probabilities (1814).* John Wiley & Sons, New York, 1902. pages: 217, 220

[65] Richard Dedekind. *Was sind und was sollen die Zahlen? (What are numbers, and what should they be?).* Vieweg, Braunschweig, 1888. pages: 426, 438

[66] Heather Demarest. *Justifying the Lebesque Measure.* Http://faculty-staff.ou.edu/D/Heather.Demarest-1/Demarest%20JLM.pdf, Cited 4 Oct. 2013, 2008. pages: 142

[67] Rene Descartes. *Discourse on the Method of rightly conducting one's reason and seeking truth in the sciences.* Http://www.earlymoderntexts.com/pdfs/descartes1637.pdf, Cited 14 Jan. 2014, 1637. pages: 1, 36

[68] Rene Descartes. *Meditations on First Philosophy in which are demonstrated the existence of God and the distinction between the human soul and body.* Http://earlymoderntexts.com/pdfs/descartes1641.pdf, Cited 18 Dec. 2013, 1637. pages: 1, 36

[69] Rene Descartes. *Principles of Philosophy.* Http://earlymoderntexts.com/pdfs/descartes1641.pdf, Cited 18 Dec. 2013, 1644. pages: 1, 36, 260

[70] David Deutsch. Quantum theory as a universal physical theory. *International Journal of Theoretical Physics*, 24(1):1–41, 1985. pages: 129

[71] David Deutsch. Quantum Theory of Probability and Decisions. *Proceedings of the Royal Society of London A*, 455(1988):3129–3137, 1999. pages: 19, 146, 170

[72] Paul A.M. Dirac. *The Principles of Quantum Mechanics*. The International Series of Monographs on Physics. Oxford University Press, 4th edition, 1966. pages: 68

[73] Freeman Dyson. Time without end: Physics and biology in an open universe. *Reviews of Modern Physics*, 51(3), 1979. pages: 328

[74] A Einstein, B Podolsky, and N Rosen. Can Quantum-Mechanical Description of Physical Reality Be Considered Complete? *Physical Review*, 47(1):777–780, 1935. pages: 103

[75] Adam Elga. Self-locating belief and the Sleeping Beauty problem. *Analysis*, 60(2):143–147, 2000. pages: 282

[76] Adam Elga. Defeating Dr. Evil with Self-Locating Belief. *Philosophy and Phenomenological Research*, 69(2):383–396, 2004. pages: 285

[77] Robert Leslie Ellis. On the Foundations of the Theory of Probability. In *The Mathematical and Other Writings of Robert Leslie Ellis*. Deighton, Bell, and Co. (Original pub.: *Transactions of the Cambridge Philosophical Society* 8, 1844), London, 1863. pages: 225

[78] Euclid. *The Elements*. Http://farside.ph.utexas.edu/euclid.html, Cited 18 Dec. 2013, 300 BC. pages: 66, 86, 148

[79] Hugh Everett III. 'Relative state' formulation of quantum mechanics. *Reviews of Modern Physics*, 29:454–462, 1957. pages: 3, 11, 12, 18, 20, 122, 126, 145, 153

[80] Hugh Everett III. The theory of the universal wave function. In *The Many-Worlds Interpretation of Quantum Mechanics*, pages 3–140. Princeton University Press, Princeton, 1973. pages: 11, 12, 13, 119, 122

[81] Hugh Everett III. *The Amoeba Metaphor*. Http://www.pbs.org/wgbh/nova/manyworlds/orig-01a.html, Cited 14 Jan. 2014, unpublished draft of 1957 dissertation, NOVA Homepage: Parallel Worlds, Parallel Lives Peter Byrne (ed.), 2008. pages: 17

[82] Hugh Everett III. *The Everett Interpretation of Quantum Mechanics*. Collected Works 1955-1980. Princeton University Press, Princeton, 2012. pages: 16, 126

[83] E Farhi, J Goldstone, and S Gutmann. How probability arises in quantum mechanics. *Annals of Physics*, 192:368–382, 1989. pages: 145, 165

[84] J Finkelstein. Has the Born rule been proven? *arXiv:0907.2064*, 2009. pages: 19

[85] Sara Foster and Harvey Brown. On a recent attempt to define the interpretation basis in the many worlds interpretation of quantum mechanics. *International Journal of Theoretical Physics*, 27(12):1507–1531, 2004. pages: 129

[86] R V Freyvald. Functions and functionals computable in the limit. *Transactions of Latvijas Vlasts Univ. Zinatn. Raksti*, 210:6–19, 1977. pages: 439, 446

[87] Travis Garrett. The theory of statistical metaphysics. In *Transvision 2004 Conference*. Trinity College, Univeristy of Toronto, Http://www.physics.unc.edu/~tmgarret/statmeta.pdf, 5-8 May, 2004. pages: 327, 405

[88] M Gell-Mann and James B Hartle. Quantum Mechanics in the Light of Quantum Cosmology. *Complexity, Entropy and the Physics of Information*, W.H. Zurek (ed.):425–459, 1990. pages: 17

[89] G Ghirardi, A Rimini, and T Weber. Unified dynamics for microscopic and macroscopic systems. *Physical Review D*, 1986. pages: 140

[90] Richard D. Gill. On an Argument of David Deutsch. *arXiv:quant-ph*, 0307188, 2004. pages: 19

[91] A Gleason. Measures on the closed subspaces of a Hilbert space. *Journal of Mathematics and Mechanics*, 6(6):885–893, 1957. pages: 2, 18, 20, 146, 197, 199

[92] Kurt Gödel. Is mathematics syntax of language. In *Kurt Gödel Collected Works Volume III*, pages 324–362. Oxford University Press, Oxford, 1986. pages: 451

[93] Kurt Gödel. *On Formally Undecidable Propositions of Principia Mathematica and Related Systems I (1931)*. Dover, Mineola, 1992. pages: 33, 174, 181, 252, 450, 451

[94] E Gold. Limiting recursion. *The Journal of Symbolic Logic*, 30(1):28–48, 1965. pages: 257, 439, 441

[95] Neill Graham. The measure of relative frequency. In *The Many-Worlds Interpretation of Quantum Mechanics*, pages 229–253. Princeton University Press, Princeton, 1973. pages: 20, 145

[96] John Graham-Cummings. The 100-year leap. *Plan 28*, Http://plan28.org, Cited 2 June 2013. pages: 254

[97] Hilary Greaves. Probability in the Everett interpretation. *Philosophy Compass*, 2(1):109–128, 2007. pages: 23

[98] R Griffiths. Consistent histories and the interpretation of quantum mechanics. *Journal of Statistical Physics*, 1984. pages: 349

[99] R Griffiths. *Consistent Quantum Theory*. Cambridge University Press, Cambridge, 2002. pages: 60, 75

[100] R Griffiths and Roland Omnes. Consistent Histories and Quantum Measurements. *Physics Today*, 52(8):26–31, 1999. pages: 17

[101] S Gutmann. Using classical probability to guarantee properties of infinite quantum sequences. *Physical Review A*, 52(5):3560–3562, 1995. pages: 145

[102] Robin Hanson. When worlds collide: quantum probability from observer selection? *Foundations of Physics*, 33(7):1129–1150, 2003. pages: 36, 145, 166, 167, 168, 169, 205, 330, 358

[103] James B Hartle. Quantum mechanics of individual systems. *American Journal of Physics*, 36(8):704–712, 1968. pages: 3, 145, 161, 162

[104] James B Hartle. The Quantum Mechanics of Closed Systems. *arXiv:quant-ph*, 9210006, 1992. pages: 349

[105] Nick Herbert. *Quantum Reality*. Anchor Books, Garden City, New York, 1985. pages: 60

[106] Jacques Herbrand. Sur la non-contradiction de l'arithmetique. *Journal fur die reine und angewandte Mathematik*, 166:1–8, 1932. pages: 252, 427

[107] David Hilbert. Die Grundlegung der elementaren Zahlenlehre. *Math. Ann.*, 104(1):485–494, 1931. pages: 440

[108] E Joos and H D Zeh. The emergence of classical properties through interaction with the environment. *Zeitschrift für Physik B Condensed Matter*, 59(2):223–243, 1985. pages: 109

[109] Immanuel Kant. *Critique of Pure Reason*. Http://earlymoderntexts.com/authors/kant.html, Cited 18 Dec. 2013, 1787. pages: 37, 49

[110] Adrian Kent. Against many worlds interpretations. *Int. J. Mod. Phys. A*, 5(9):1745–1762, 1990. pages: 125

[111] Adrian Kent. Everett and Evolution. In *Many Worlds at 50 (Perimeter Institute)*, September 2007. pages: 13

[112] John Maynard Keynes. *A Treatise on Probability*. MacMillan and Co., London, 1909. pages: 222

[113] Stephen C Kleene. A theory of positive integers in formal logic. Part II. *American Journal of Mathematics*, 57:219–244, 1935. pages: 252

[114] Stephen C Kleene. General recursive functions of natural numbers. *Math. Annalen*, 112:727–742, 1936. pages: 252

[115] Stephen C Kleene. On notation for ordinal numbers. *Journal of Symbolic Logic*, 3:150–155, 1938. pages: 264

[116] Stephen C Kleene. Recursive predicates and quantifiers. *Trans. Amer. Math. Soc.*, 53:41–73, 1943. pages: 252, 264, 441, 442, 443, 444, 450

[117] Stephen C Kleene and J Rosser. The inconsistency of certain formal logics. *Annals of Mathematics*, 1935. pages: 431

[118] S Kochen and E Specker. The problem of hidden variables in quantum mechanics. *Journal of Mathematics and Mechanics*, 17(1):59–87, 1967. pages: 198

[119] A. N. Kolmogorov. Three approaches to the quantitative definition of information. *International Journal of Computer Mathematics*, 2(1):157–168, 1968. pages: 243

[120] Julius König. *Neue Grundlagen der Logik, Arithmetik und Mengenlehre*. Veit, Leipzig, 1914. pages: 255, 431

[121] Leon Gordon Kraft. *A device for quantizing, grouping, and coding amplitude-modulated pulses*. PhD thesis, Massachusetts Institute of Technology, 1949. pages: 247

[122] Saul A Kripke. *Naming and Necessity*. Wiley-Blackwell, Hoboken, 1991. pages: 37

[123] Gottfried Wilhelm Leibniz. *The Principles of Philosophy known as Monadology*. Http://earlymoderntexts.com/pdfs/leibniz1714b.pdf, Cited 18 Dec. 2013, 1720. pages: 1

[124] Gottfried Wilhelm Leibniz. Estimating the Uncertain (1678). In M Dascal, editor, *The Art of Controversies*, pages 105–118. Springer, 2006. pages: 217, 220

[125] Matthew Leifer. Anyone for frequentist fudge? *Quantum Quandaries*, Http://mattleifer.wordpress.com/2006/06/14/anyone-for-frequentist-fudge, Cited 20 Aug. 2011, June 2006. pages: 160

[126] John Leslie. A difficulty for Everett's many-worlds theory. *International Studies in the Philosophy of Science*, 10(3):239–246, 1996. pages: 341

[127] David Lewis. Attitudes De Dicto and De Se. *The Philosophical Review*, 88(4):513–543, 1979. pages: 288

[128] David Lewis. A subjectivist's guide to objective chance. University of California Press, 1980. pages: 180

[129] David Lewis. *On the Plurality of Worlds*. Wiley-Blackwell, Hoboken, 1986. pages: 283

[130] David Lewis. Sleeping Beauty. *Analysis*, 61(3):171–187, 2001. pages: 282

[131] David Lewis. How Many Lives Has Schrödinger's Cat? *Australasian Journal of Philosophy*, 82(1):3–22, 2004. pages: 341

[132] Peter J Lewis. Deutsch on quantum decision theory. *PhilSci Archive*, philsci-archive.pitt.edu, 2003. pages: 19

[133] Peter J Lewis. Quantum Sleeping Beauty. *Analysis*, 67(293):59–65, 2007. pages: 323

[134] Ming Li and P M B Vitányi. *An introduction to Kolmogorov complexity and its applications*. Springer, New York, 2008. pages: 247, 248, 254

[135] Fabrizio Logiurato and Augusto Smerzi. Born Rule and Noncontextual Probability. *arXiv:quant-ph*, 1202.2728, 2012. pages: 139, 198, 201, 390

[136] Paul Lorenzen. lgebraische und logistische Untersuchungen über freie verbände. *Journal of Symbolic Logic*, 16:81–106, 1951. pages: 440

[137] J R Lucas. Minds, Machines, and Gödel. *Philosophy*, 36:112, 1961. pages: 262

[138] Jacques Mallah. Many-worlds interpretations can not imply 'quantum immortality'. *arXiv:quant-ph*, 0902.0187, 2006. pages: 205, 332, 344

[139] Jacques Mallah. The Many Computations Interpretation (MCI) of Quantum Mechanics. *arXiv:quant-ph*, 0709.0544, 2007. pages: 145, 327, 405

[140] R L McFarland, W T Powers, and R K Clark. A general feedback theory of human behavior: A prospectus. *The American Psychologist*, 12(7):462, 1957. pages: 402

[141] Norman Megill. Metamath Proof Explorer. *us.metamath.org*, Http://us.metamath.org/mpegif/mmset.html, Cited 3 Oct. 2013. pages: 434

[142] D H Mellor. *Probability: a philosophical introduction*. Routledge, Abingdon, 2005. pages: 212, 225, 239

[143] Luigi Federico Menabrea and Countess of Lovelace Ada Augusta. Sketch of the Analytical Engine invented by Charles Babbage. *Bibliothèque Universelle de Genève*, 82, 1842. pages: 252, 254

[144] Arnold Neumaier. *On the many-worlds interpretation*. Http://www.mat.univie.ac.at/~neum/manyworlds.txt, Cited 9 Oct. 2008, 1999. pages: 125

[145] R Omnes. A New Interpretation of Quantum Mechanics and Its Consequences in Epistemology. *Foundations of Physics*, 25(4):605–629, 1995. pages: 17

[146] Hans P Moravec. *Mind Children: the future of robot and human intelligence*. Harvard University Press, Cambridge, 1988. pages: 205, 332, 340

[147] Don N. Page. Sensible Quantum Mechanics: Are Only Perceptions Probabilistic? *arXiv:quant-ph*, 9506010, 1995. pages: 328

[148] Stephen Palmquist. *Kant's System of Perspectives: an architectonic interpretation of the critical philosophy*. University Press of America, Lanham, 1993. pages: 37

[149] Parmenides. *On Nature*. Http://allanrandall.ca/Parmenides.html, Cited 19 Oct. 2013, c. 475 BC. pages: 1, 327, 405

[150] Vern Paulsen, B Bollobás, W Fulton, A Katok, F Kirwan, and P Sarnak. *Completely Bounded Maps and Operator Algebras*. Cambridge University Press, Cambridge, 2003. pages: 116

[151] Charles Sanders Peirce. Notes on the doctrine of chances. In *Collected Papers of Charles Sanders Peirce*, pages 661–668. Harvard University Press, Cambridge, 1958. pages: 228

[152] Roger Penrose. Precis of The Emperor's New Mind: Concerning computers, minds, and the laws of physics. *Behavioral and Brain Sciences*, 13:643–705, 1989. pages: 137

[153] Roger Penrose. *The Emperor's New Mind: concerning computers, minds, and the laws of physics*. Oxford University Press, Oxford, 1989. pages: 137, 262

[154] Roger Penrose. *Shadows of the Mind: a search for the missing science of consciousness*. Oxford University Press, Oxford, 1994. pages: 137

[155] R M Perry. *Forever For All*. Universal Publishers, Boca Raton, 2000. pages: 205, 332, 340

[156] Rózsa Péter. Konstruktion nichtrekursiver Funktionen. *Mathematische Annalen*, 111:42–60, 1935. pages: 427, 438

[157] J C Polkighorne. *The Quantum World*. Princeton University Press, Princeton, 1984. pages: 17, 332

[158] Karl Popper. *The Logic of Scientific Discovery*. Routledge, London, 1959. pages: 306, 307, 380

[159] Karl Popper. The propensity interpretation of probability. *The British Journal for the Philosophy of Science*, 10(37):25–42, 1959. pages: 225, 228

[160] Emil L Post. Finite combinatory processes. *Journal of Symbolic Logic*, 1:103–105, 1936. pages: 252

[161] William T Powers. *Behavior: The control of perception*. Hawthorne, New York, 1973. pages: 393, 402

[162] William T Powers, R K Clark, and R L McFarland. A general feedback theory of human behavior: Part I. *Perceptual and Motor Skills*, 11(1)(1):71–88, 1960. pages: 402

[163] William T Powers, R K Clark, and R L McFarland. A general feedback theory of human behavior: Part II. *Perceptual and Motor Skills*, 11(3)(3):309–323, 1960. pages: 402

[164] Huw Price. Decisions, Decisions, Decisions: Can Savage Salvage Everettian Probability? *arXiv:quant-ph*, 0802.1390, 2008. pages: 19

[165] Hilary Putnam. Trial and error predicates and the solution to a problem of Mostowski. *Journal of Symbolic Logic*, 30(1):49–57, 1965. pages: 257, 439, 441, 445

[166] Willard Van Orman Quine. *Ontological Relativity and other essays*. Columbia University Press, New York, 1969. pages: 268

[167] F P Ramsey. Truth and probability (1926). In *The Foundations of Mathematics and Other Logical Essays*, pages 156–198. Harcourt, Brace and Company, New York, 1931. pages: 239

[168] Allan F Randall. *Computational Platonism*. Http://allanrandall.ca/Plato.html, Cited 6 May 2013, 1995. pages: 405

[169] Allan F Randall. *Quantum superposition, necessity and the identity of indiscernibles*. Http://allanrandall.ca/Indiscernibles.html, Cited 7 Oct. 2013, 1996. pages: 4, 58, 372, 405

[170] Allan F Randall. *Truth, coherence and correspondence in the metaphysics of F.H. Bradley*. Http://allanrandall.ca/Bradley.html, Cited 6 May 2013, 1996. pages: 283, 405

[171] Allan F Randall. *Logic, idealism and materialism in early and late Wittgenstein*. Http://allanrandall.ca/Wittgenstein.html, Cited 6 May 2013, 1997. pages: 405

[172] Allan F Randall. *Quantum phenomenology*. Http://allanrandall.ca/Phenomenology.html, Cited 6 May 2013, Toronto, 1997. pages: 4, 58, 205, 372, 405

[173] Allan F Randall. *A critique of the Kantian view of geometry*. Http://allanrandall.ca/Geometry.html, Cited 6 May 2013, 1998. pages: 39

[174] Allan F Randall. Quantum miracles and immortality. In *Transvision 2004 Conference*. Trinity College, Univeristy of Toronto, Http://allanrandall.ca/tv2004.pdf, 5-8 May, 2004. pages: 4, 36, 57, 58, 168, 205, 329, 332, 341, 344, 372, 405

[175] Allan F Randall. Chapter 15. In *Meet the philosophers of ancient Greece*. Ashgate, 2005. pages: 1

[176] Allan F Randall. *Limit recursion and Gödel's incompleteness theorem*. MA thesis. York University, 2006. pages: 174, 181, 251, 451, 453

[177] Hartley Rogers. Theory of recursive functions and effective computability. McGraw-Hill Book Company, New York, 1967. pages: 254

[178] Hartley Rogers, Jr. Gödel numberings of partial recursive functions. *The Journal of Symbolic Logic*, 23(3):331–341, 1958. pages: 267, 269

[179] S Roman. *Introduction to coding and information theory*. Springer Verlag, 1997. pages: 247

[180] J B Rosser. Gödel theorems for non-constructive logics. *Journal of Symbolic Logic*, 2:129–137, 1937. pages: 440

[181] Antoine Royer. Antilinear operators in Dirac's bra-ket notation. *American Journal of Physics*, 62(8):730–732, 1994. pages: 73

[182] M A Rubin. Relative Frequency and Probability in the Everett Interpretation of Heisenberg-Picture Quantum Mechanics. *Foundations of Physics*, 33(3):379–405, 2003. pages: 145

[183] Simon Saunders. Decoherence, relative states, and evolutionary adaptation. *Foundations of Physics*, 23(12):1553–1585, 1993. pages: 129

[184] Simon Saunders. Time, quantum mechanics, and decoherence. *Synthèse*, 102(2):235–266, 1995. pages: 129

[185] Simon Saunders. Time, Quantum Mechanics, and Probability. *Synthèse*, 114:373–404, 1998. pages: 141, 142

[186] Leonard J Savage. *The foundations of statistics*. Dover, 1972. pages: 170, 239

[187] Maximilian Schlosshauer. Experimental motivation and empirical consistency in minimal no-collapse quantum mechanics. *Annals of Physics*, 321(1):112–149, 2006. pages: 137

[188] Maximilian Schlosshauer and Arthur Fine. On Zurek's derivation of the Born rule. *arXiv:quant-ph*, 0312058, 2003. pages: 191, 194

[189] Maximilian A. Schlosshauer. *Decoherence and the quantum-to-classical transition*. Springer, Berlin Heidelberg New York, 2007. pages: 60, 109, 113

[190] Juergen Schmidhuber. A Computer Scientist's View of Life, the Universe, and Everything. *arXiv:quant-ph*, 9904050, 1999. pages: 327

[191] Juergen Schmidhuber. Algorithmic Theories of Everything. *arXiv:quant-ph*, 0011122, 2000. pages: 327, 405

[192] Erhard Schmidt. Zur Theorie der linearen und nichtlinearen Integralgleichungen. *Math. Ann.*, 63:433–476, 1907. pages: 102

[193] Moses Schönfinkel. On the building blocks of mathematical logic. In *From Frege to Gödel*, pages 355–366. Harvard University Press, Cambridge, 1967. pages: 252, 255, 258, 431, 439

[194] E Schrödinger and J Trimmer. The Present Situation in Quantum Mechanics: A Translation of Schrödinger's "Cat Paradox" Paper. *Proceedings of the American Philosophical Society*, 124(5):323–338, 1980. pages: 8, 106

[195] B Schumacher. Quantum coding. *Physical Review A*, 51(4):2738–2747, 1995. pages: 119, 120

[196] Kurt Schütte. Beweistheoretische erfassung der unendlichen induktion in der zahlentheorie. *Math. Ann.*, 122:369–389, 1951. pages: 440

[197] Peter Sestoft. Lambda calculus reduction workbench. *www.itu.dk*, 1995. pages: 431

[198] R Shankar. *Principles of Quantum Mechanics (2nd Edition)*. Springer, Berlin, Heidelberg, 1994. pages: 60, 64, 80

[199] Claude E Shannon. A mathematical theory of communication. *Bell Sys. Tech. J.*, 27:379–423,623–656, 1948. pages: 118, 242

[200] Himanshu Sharma and R Srikanth. No-signaling from Gleason non-contextuality and the tensor-product structure. *arXiv:quant-ph*, 1202.1804, 2012. pages: 203

[201] Lee Smolin. Did the Universe evolve? *Classical and Quantum Gravity*, 9:173–191, 1992. pages: 306

[202] Lee Smolin. Scientific alternatives to the anthropic principle. In *Universe Or Multiverse?*, pages 323–266. Cambridge University Press, June 2007. pages: 306, 309, 310, 311, 313

[203] Ray J Solomonoff. *A Preliminary Report on a General Theory of Inductive Inference*. Zator Company and United States Air Force Office of Scientific Research, ZTB-138, 1960. pages: 56, 243

[204] Ray J Solomonoff. A formal theory of inductive inference: parts 1 and 2. *Information and Control*, 7(1-2):1–22 & 224–254, 1964. pages: 56, 243, 276, 280

[205] Ray J Solomonoff. Complexity-based induction systems: comparisons and convergence theorems. *IEEE Transactions on Information Theory*, 24(4):422–432, 1978. pages: 56, 274

[206] Ray J Solomonoff. The discovery of algorithmic probability. *Journal of Computer and System Sciences*, 55, 1997. pages: 277

[207] Baruch Spinoza. *Ethics Demonstrated in Geometrical Order*. Http://earlymoderntexts.com/pdfs/spinoza1665.pdf, Cited 18 Dec. 2013, 1677. pages: 1

[208] Henry P Stapp. The basis problem in many-worlds theories. *Canadian Journal of Physics*, 80(9):1043–1052, 2002. pages: 129, 130

[209] G Strang. *Introduction to Linear Algebra*. Wellesley-Cambridge Press, Wellesley, 2003. pages: 380

[210] David Strayhorn. Derivation of the Born rule from outcome counting and a solution to the quantitative problem of the multiple worlds interpretation. *Philica.com*, 28, 2006. pages: 145

[211] Max Tegmark. Importance of quantum decoherence in brain processes. *Physical review. E, Statistical physics, plasmas, fluids, and related interdisciplinary topics*, 61(4 Pt B):4194–4206, April 2000. pages: 114, 135, 137

[212] Max Tegmark. The mathematical universe. *Foundations of Physics*, 38(2):101–150, 2008. pages: 327, 329, 404, 405

[213] Max Tegmark. Many Worlds in Context. *arXiv:quant-ph*, 0905.2182, 2009. pages: 140

[214] Max Tegmark. *Our Mathematical Universe: My Quest for the Ultimate Nature of Reality*. Knopf, New York, 2014. pages: 205, 327, 344

[215] Frank J Tipler. *The Physics of Immortality: modern cosmology, God, and the resurrection of the dead*. Random House, New York, 1994. pages: 205, 340

[216] John Tromp. *John's Combinatory Logic Playground*. Http://homepages.cwi.nl/~tromp/cl/cl.html, Cited 6 Apr. 2013, 1999. pages: 432

[217] Alan Turing. On computable numbers with an application to the Entscheidungsproblem. *Proc of the London Mathematical Society, Series 2*, 42-43:230–265, 544–546, 1936. pages: 250, 252, 266, 276, 327, 360, 427, 428, 430, 438

[218] Alan M Turing. Computability and λ-definability. *Journal of Symbolic Logic*, 2(4):153–163, 1937. pages: 250, 252, 253, 427, 430, 432

[219] Alan M Turing. Systems of logic based on ordinals. *Proc. London Mathematical Society*, 45:161–228, 1939. pages: 440

[220] Alan M Turing. *Intelligent Machinery*. National Physical Laboratory, 1948. pages: 250, 427

[221] John Venn. *The Logic of Chance: An Essay on the Foundations and Province of the Theory of Probability, with especial reference to its logical reasons and its application to moral and social science*. MacMillan and Co., London, 1876. pages: 225

[222] Richard Von Mises. *Probability, Statistics and Truth*. Dover, 2nd revised edition, 1981. pages: 225, 226, 227

[223] John von Neumann. *Mathematical Foundations of Quantum Mechanics (*Mathematische Grundlagen der Quantenmechanik, 1932). Princeton University Press, Princeton, 1955. pages: 118, 122, 201

[224] John Walker. Java Applet Analytical Engine Emulator. *fourmilab.ch*, Http://www.fourmilab.ch/babbage/applet.html, Cited 3 June 2007. pages: 254

[225] David Wallace. Quantum probability from subjective likelihood: Improving on Deutsch's proof of the probability rule. *Studies in History and Philosophy of Modern Physics*, 38:311–332, 2007. pages: 19, 170, 176, 187

[226] David Wallace. A formal proof of the Born rule from decision-theoretic assumptions. *arXiv:quant-ph*, 0906.2718, 2009. pages: 19, 141, 146, 187, 198, 201

[227] David Wallace. How to prove the Born rule. In *Many Worlds? Everett, quantum theory, and reality*, pages 227–263. Oxford University Press, Oxford, 2010. pages: 22, 134, 170, 171, 180, 181, 187, 188, 366

[228] Hao Wang. The formalization of mathematics. *Journal of Symbolic Logic*, pages 241–266, 1954. pages: 440

[229] M B Weissman. Emergent measure-dependent probabilities from modified quantum dynamics without state-vector reduction. *Foundations of Physics Letters*, 12(5):407–426, 1999. pages: 145

[230] David Wick. *The Infamous Boundary: Seven Decades of Heresy in Quantum Physics*. Copernicus, New York, 1996. pages: 17

[231] Norbert Wiener. *Cybernetics: or Control and Communication in the Animal and the Machine*. Cambridge, 1948. pages: 393, 402

[232] Eugene Wigner. *Group Theory and its Application to the Quantum Mechanics of Atomic Spectra*. Academic Press, New York, London, 1959. pages: 82, 378

[233] Eugene Wigner. Remarks on the mind-body question. In *The Scientist Speculates*. William Heinemann, London, 1961. pages: 7

[234] Ernst Zermelo. Uber stufen der quantifikation und die logik des unendlichen. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 41:85–88, 1931. pages: 440

[235] Ernst Zermelo. Grundlagen einer allgemeinen theorie der mathematischen satzsysteme. *Fundamenta Mathematicae*, 25:136–146, 1935. pages: 440

[236] Wojciech Hubert Zurek. Decoherence and the transition from quantum to classical. *Physics Today*, 44(10):36, 1991. pages: 109, 117

[237] Wojciech Hubert Zurek. Quantum Darwinism and Envariance. *arXiv:quant-ph*, 0308163, 2003. pages: 191, 194

[238] Wojciech Hubert Zurek. Probabilities from Entanglement, Born's Rule from Envariance. *arXiv.org*, May 2004. pages: 114, 191, 202

[239] Wojciech Hubert Zurek. Relative States and the Environment: Einselection, Envariance, Quantum Darwinism, and the Existential Interpretation. *arXiv:quant-ph*, 0707.2832, 2007. pages: 191

[240] Wojciech Hubert Zurek. Quantum Darwinism. *arXiv:quant-ph*, 0903.5082, 2009. pages: 191

A More Schrödinger's Equations

While the very general forms of Schrödinger's equation in Ch. 2, expressed in natural units, were well-suited for our purposes, some readers might be confused about the relationship between these simple versions and those found in many textbooks, which are more complicated and usually written in synthetic (non-natural) units. To address these concerns, this appendix provides some brief derivations of some of the more common specialized forms of the dynamical Schrödinger's equation. (While this material can be found in any number of introductory textbooks, it is presented here in a form that follows naturally from the development in the main text.)

A.1 Specific time-independent Schrödinger equations

From the most general version of the time-independent Schrödinger's equation, already derived in the main text as (2.124):

$$\frac{\partial}{\partial x}|\psi\rangle = i\hat{H}|\psi\rangle \tag{A.1}$$

we can calculate any number of more specific forms, by deriving specific versions of the Hamiltonian depending on the situation. Some examples follow.

A.1.1 Stationary states and atomic orbitals

The basic case (also derived in the main text as (2.123)) has a Hamiltonian simply equal to the spatial frequency:

$$\begin{aligned} \hat{H} &= k \\ \hat{H}|\psi\rangle &= k|\psi\rangle \\ \frac{\partial}{\partial x}|\psi\rangle &= ik|\psi\rangle \end{aligned} \tag{A.2}$$

This occurs when the wavefunction $|\psi\rangle$ is a time-independent "stationary state" and the only spatial dependence is rotation through the complex plane, as is the case for the standing waves that characterize atomic orbitals ((2.125) in the main text) for which the eigenvalues correspond to the wavenumbers ($k = 1, 2, \ldots$) and energy levels E_k.

$$\hat{H}|\psi\rangle = E_k|\psi\rangle \tag{A.3}$$

A.1.2 Mechanical energy

For a great many practical applications, we can define the Hamiltonian in terms of total mechanical energy (E), as a sum of kinetic (E_k) and potential (E_P) energies:

$$E = E_k + E_p \tag{A.4}$$

Kinetic energy (E_k) is usually expressed as

$$E_k = \frac{1}{2}mv^2 = \frac{p^2}{2m} \tag{A.5}$$

but we are not yet taking differing masses into account, so for us $m = 1$, and momentum is simply the wavenumber,

$$p = k \tag{A.6}$$

and

$$E_K = f = \frac{1}{2}k^2 \tag{A.7}$$

giving us a Hamiltonian of

$$\hat{H} = \frac{1}{2}k^2 + E_p \tag{A.8}$$

which yields

$$\hat{H}\left|\psi\right\rangle \quad = \quad \frac{1}{2}k^2\left|\psi\right\rangle + E_p\left|\psi\right\rangle \tag{A.9}$$

We can now differentiate Schrödinger's equation in terms of x (this is the second differentiation of the wave equation with respect to x, since it was already differentiated in order to obtain Schrödinger's equation in the first place):

$$\frac{\partial^2}{\partial x^2}\left|\psi\right\rangle = (ik)^2\left|\psi\right\rangle \tag{A.10}$$

Substituting into (2.123),

$$-\frac{\partial^2}{\partial x^2}\left|\psi\right\rangle \quad = \quad k^2\left|\psi\right\rangle \tag{A.11}$$

Substituting for $\dot{k}^2\left|\psi\right\rangle$ from (A.9):

$$\hat{H}\left|\psi\right\rangle = -\frac{1}{2}\frac{\partial^2}{\partial x^2}\left|\psi\right\rangle + E_p\left|\psi\right\rangle \tag{A.12}$$

which is our more specific form of the time-independent Schrödinger's equation.

A.1.3 Mechanical energy (with mass)

To take mass m into account we use $E_k = {}^{k^2}/_{2m}$ and we get

$$\hat{H} = \frac{k^2}{2m} + E_p \tag{A.13}$$

which gives us

$$\hat{H}\left|\psi\right\rangle = -\frac{1}{2m}\frac{\partial^2}{\partial x^2}\left|\psi\right\rangle + E_p\left|\psi\right\rangle \tag{A.14}$$

A.1.4 Non-natural energy units

In practical applications, it is often helpful to convert from the natural units we have been working in (where energy is dimensionless and momentum is simply the wavenumber $p = k$) to dimensioned energy units (where the unit of energy is Planck's constant h, so that $p = hk$). In natural units, h is not needed, since we just declare that $h = 1$. Our last equation above now becomes

$$\hat{H}\,|\psi\rangle = -\frac{h^2}{2m}\frac{\partial^2}{\partial x^2}\,|\psi\rangle + E_p\,|\psi\rangle \tag{A.15}$$

This is really just the same formula; it simply assigns a naturally dimensionless quantity (k) a synthetic "unit" h. In natural units, we set $h = 1$, and its inclusion in equations is therefore optional.

This is still not *quite* the common form of the equation found in textbooks, however, since it is actually more common to use angular spatial frequency $\dot{k}$ and the *reduced* Planck's constant ($\hbar = {}^{h}/_{\tau}$) to derive these equations, instead of k and h, as we have done here. In that case, one would prefer to say (equivalently) that $\hbar = 1$ and that $p = \hbar\dot{k}$. In ASU terms, it makes more sense to use k and h, since we are working in a fundamentally discrete domain and k simply takes on the values of the natural numbers (it is our "counter" or "iterator"). In common practice, however, $\dot{k}$ (which is more commonly just called k) and $\hbar$ are more popular choices, resulting in the exact same formula, except with the reduced version of Planck's constant:

$$\hat{H}\,|\psi\rangle = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\,|\psi\rangle + E_p\,|\psi\rangle \tag{A.16}$$

There is no formal difference between our three versions of this same equation. The difference is only a matter of our choice of units. Including h or $\hbar$ at all, instead of just using 1, is an attempt to facilitate ready conversion to other (synthetic or non-natural) systems of units that are a better match with the scale of human activities. For instance, to convert to SI units, we can use the following conversion factor:

$$\hbar \approx 1.054571726 \times 10^{-34}\mathrm{J}\cdot\mathrm{s} \tag{A.17}$$

A.2 Specific time-dependent Schrödinger equations

From the most general version of the time-dependent Schrödinger's equation, already derived in the main text as (2.128),

$$i\frac{\partial}{\partial t}\,|\psi\rangle \quad = \quad \hat{H}\,|\psi\rangle \tag{A.18}$$

we can calculate any number of more specific forms, by deriving specific versions of the Hamiltonian depending on the situation. Some examples follow.

A.2.1 Angular frequency

The basic case, already derived in the main text as (2.127), is where the Hamiltonian is just the angular frequency,

$$\hat{H} = \dot{f} \tag{A.19}$$

giving

$$i\frac{\partial}{\partial t}\,|\psi\rangle = \dot{f}\,|\psi\rangle \tag{A.20}$$

A.2.2 Mechanical energy

Going back to the time-independent Schrödinger's equation for mechanical energy:

$$\hat{H}\left|\psi\right\rangle = -\frac{1}{2}\frac{\partial^2}{\partial x^2}\left|\psi\right\rangle + E_p\left|\psi\right\rangle \tag{A.21}$$

we substitute for $\hat{H}\psi$ from (2.128),

$$i\frac{\partial}{\partial t}\left|\psi\right\rangle = -\frac{1}{2}\frac{\partial^2}{\partial x^2}\left|\psi\right\rangle + E_p\left|\psi\right\rangle \tag{A.22}$$

which is our desired, more specific, form of the time-dependent Schrödinger's equation.

A.2.3 Mechanical energy (with mass)

To take mass m into account we, again, use $E_k = {}^{k^2}/_{2m}$ and we get

$$i\frac{\partial}{\partial t}\left|\psi\right\rangle = -\frac{1}{2m}\frac{\partial^2}{\partial x^2}\left|\psi\right\rangle + E_p\left|\psi\right\rangle \tag{A.23}$$

A.2.4 Non-natural energy units

We can, again, express this in terms of dimensioned energy units $\hbar$:

$$i\hbar\frac{\partial}{\partial t}\left|\psi\right\rangle = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\left|\psi\right\rangle + E_p\left|\psi\right\rangle \tag{A.24}$$

B Recursion Theory

B.1 Introduction

This appendix provides some further details on some of the foundational recursive languages, going beyond the brief survey given in the main text, and providing more formal descriptions of some of the languages.

B.2 Functions

Definition B.1. For our purposes, a "function" will mean a mapping from natural numbers ("input arguments") to a unique natural number ("output value" or just "value").

Definition B.2. A "predicate" is a function, interpreted as returning only one of two values, "true" or "false", where 0 is interpreted as false, and any other natural number is interpreted as true.

A function may have some number of variables, or input arguments. For the function $f(x)$, there is one input argument, x.

Definition B.3. A "function call" is the use or "application" of a function, where all input arguments are assigned determinate values, resulting in an output value for the function.

For example, $f(2)$ is a call to function $f()$ with input argument $x = 2$. While every call has a value, some values may be described as "determinate", while others may be described as "indeterminate". A common example would be the division function, where for example $\div(6, 2)$ returns the value 3, while $\div(6, 0)$ has an indeterminate value, since we cannot divide by zero, and we say the output value is undefined.

Definition B.4. A "total function" is any function that always returns a determinate value, no matter what values are assigned to its input arguments.

Definition B.5. A "partial function" is a function that may, or may not, have an undefined output value for some values of its input arguments.

I will generally assume a single input argument unless otherwise specified.

Functions whose calls can be computerized on an appropriately programmed computer, assuming unlimited memory and computing time, are the "partial computable functions" (which will be more formally defined shortly). The partial computable functions are also definable as "Turing machines" and as expressions in the "λ-calculus"'. These three languages form an equivalence class of languages, called "Turing-complete"', and together represent our most comprehensive model of computation. A "total computable function" is a partial computable function whose calls are always guaranteed to halt and return a value within a finite number of computational steps, no matter what values are bound to its input arguments.

B.3 Arithmetic

We will talk about arithmetic (arithmetical number theory) in two different forms:

1. *Peano Arithmetic*: first-order predicate logic plus the Peano axioms.
2. *Arithmetical Hierarchy*: a hierarchy of sets defined by first-order arithmetical predicates.

B.3.1 Peano arithmetic

We will define only Peano Arithmetic in this section, also called Z (see [21]). The Arithmetical Hierarchy will be defined later, in Appendix C (on limit computation). To get Finite Arithmetic (or Q), replace Z7 with $\forall x(x \neq 0 \rightarrow \exists y(x = s(y)))$. Note that x and y stand for arbitrary arithmetical expressions.

- First-order predicate expressions, with equality $(x = y)$.
- Arithmetical expressions:
  0
  $s(x)$
  $+(x, y)$
  $\times(x, y)$
- Axioms:
  Z1. $\forall x \forall y(s(x) = s(y) \rightarrow x = y)$
  Z2. $\forall x(0 \neq s(x))$
  Z3. $\forall x(x + 0 = x)$
  Z4. $\forall x \forall y(x + s(y) = s(x + y))$
  Z5. $\forall x(x \times 0 = 0)$
  Z6. $\forall x \forall y(x \times s(y) = (x \times y) + x)$
  Z7. $(F(0)\ \&\ \forall x(F(x) \rightarrow F(s(x)))) \rightarrow \forall x F(x)$
  where $F(x)$ stands for any formula of one free variable x.

B.4 Recursive Function Theory

B.4.1 Basic Functions

Before we look at the various kinds of recursive functions, we start by defining three basic functions common to all:

Definition B.6. The primitive constant function 0, usually identified with the number zero.

Definition B.7. The 1-argument successor function, $s(x)$, where input argument "x" can be the primitive constant 0, or any call to $s()$. Calls to $s()$ are generally associated with the natural numbers, and all calls to $s()$ end up bottoming out at some point and calling 0. For instance, $s(s(s(0)))$ is usually identified with the number 4 (so we will sometimes just go ahead and write this as "4", for short).

Definition B.8. The n-argument projector function, $p_i(x_1, x_2, \ldots)$, which picks out the i^{th} argument and returns it. So $p_2(3, 5, 2, 0, 9)$ returns the value 5. As written, $p_i()$ is actually a family of functions, indexed by i (although this particular family could actually be replaced with a single function, if desired, with i as the first input argument). If i is greater than the number of arguments, then the function just defaults to returning the last element. If $i = 0$, the function defaults to returning the first element.

B.4.2 Composition

Definition B.9. Given the n-argument function $f(x_1, \ldots, x_n)$, one can create the "composition" of this function with any number k of other functions $g_1(x_1, \ldots, x_k), \ldots, g_n(x_1, \ldots, x_k)$, to yield a new function, $h()$,

$$h(x_1, \ldots, x_k) = f(g_1(x1_1, \ldots, x_k), \ldots, g_n(x_1, \ldots, x_k)) \tag{B.1}$$

For example, the following defines $h(x, y) = x + y + 6$:

$$\begin{aligned} f(x) &= s(x) \\ g_1(x, y) &= s(x) + s(y) \\ g_2(x) &= s(s(x)) \end{aligned} \tag{B.2}$$

So, using composition, we can define a wide variety of functions in terms of previously defined ones. It is not permitted, however, to define a function in terms of itself or in terms of a function that has not been defined.

B.5 Primitive Recursive Functions (bounded loops)

Primitive recursion, first defined by [65], is a nontrivial system of computation, and while many interesting functions can be defined with it, it is weaker than that yielded by Turing computability.

Recursion in general allows a function to recur, or call itself. This potentially allows infinite recurrence, and the function, once called, may never halt. Primitive recursion is defined so as to avoid this possibility.

Definition B.10. "Primitive recursive functions" are those obtainable from a finite number of applications of the three basic (zero, successor and projector) functions, composition, and functions formable by "primitive recursion", meaning of the following form:

$$f(x_1, \ldots, x_n, 0) = g(\ldots) \tag{B.3}$$

$$f(x_1, \ldots, x_n, s(y)) = h(\ldots, y, f(x_1, \ldots, x_n, y)) \tag{B.4}$$

where $f()$ and $g()$ are primitive recursive.

The first line above is called the "base clause", as it happens only in the case of 0 as an input argument, and hence signals the end of the recursion (since the input argument decreases by one with each recursive call). The rest of the recursive calls consist of the second line above, called the "main clause". Note that while we may call a function within its own definition, this must be done in such a way that one of the arguments gets decremented, meaning it goes from $s(y)$ to y. This way, the function calls itself, which calls itself, and so on and so on, but since one of the arguments is getting decremented by 1 each time, that argument will eventually reach 0. At that point, the base clause is invoked, telling the function what value to return now that the decremented argument has reached 0. Since the decremented argument, sometimes called a "counter", is guaranteed to reach 0 eventually, and the base clause itself is known to be primitive recursive, the function call is guaranteed to terminate.

B.6 Partial and total computable functions (unbounded loops)

In addition to the guarantee of termination, primitive recursive functions are restricted by requiring that, in the application of (equation (B.4)), it must be possible to define $f()$ such that it calls itself only once in its own definition. Otherwise, fancier, more complex types of functions are possible that are not considered primitive recursive. If we *also* remove the restriction that loops be bounded, we get the "partial recursive" functions. If, while dropping the restriction that loops be bounded in their definition, we still require that all loops eventually halt anyway, we get the "total computable functions". Because they always halt, they are total, like the primitive recursive functions.

Gödel effectively used total computable functions in the proof of his incompleteness theorem, although his work lacks a general definition of them. The notion was later clarified by [106].

Definition B.11. We define the μ, or "unbounded minimization" operator, so that given a partial function $f(x_1, \ldots, x_n, y)$, the partial function $\mu_y f()$ is defined such that:

$$\mu_y f(x_1, \ldots, x_n) = \begin{cases} \text{the minimum value for } y \text{ for which } f(x_1, \ldots, x_n) = 0, \text{ if such a value exists,} \\ \textit{undefined} \text{ otherwise.} \end{cases} \tag{B.5}$$

The phrase $x_1, \ldots, x_n$ will be simplified to x (i.e., $n = 1$) for the purposes of informal discussion, since the precise number of inputs arguments is not of interest. The unbounded minimization version of a function returns the minimum value of y that makes $f(x, y) = 0$.

Definition B.12. "Partial recursion functions" are those obtainable from a finite number of applications of the three basic (zero, successor and projector) functions, composition, primitive recursion and the μ operator.

Definition B.13. "Total computable functions" are the total partial computable functions.

Partial computable functions differ from primitive recursive functions because they are not necessarily total. The computable functions are just those partial computable functions that are total. While it is not immediately obvious that the computable functions are not simply equivalent to the primitive recursive functions, as it turns out, there are functions that can be shown to be computable but *not* primitive recursive, for instance the Ackermann function [1, 156]. The partial computable functions can be shown to be equivalent to those functions that can be formulated as Turing machines, and thus, they are often taken as a general model of computation.

Definition B.14. A "computation" is the running of a particular Turing machine, or equivalently, the calculation of the value of any given application of (call to) a partial computable function.

Partial computable functions are thus essentially equivalent to the use of unbounded loops in computer programming, such as the `WHILE` loops in `BASIC` or related languages. Total computable functions are essentially the same, with the restriction that, whatever inputs are given to a program, it must always halt, making them essentially the same as the computational notion of an "algorithm" or "effective procedure". At least, it is *widely believed* that total computable functions completely formalize the intuitive notion of a mechanical effective procedure or algorithm, although such has never been proved. This conjecture is known as the "Church-Turing Thesis" [56, 217, 220, 218].

Note that no real-world computer programming language could be considered a scheme for constructing total computable functions, since there exists no constructive definition of the total computable functions, as opposed to the partial ones. General recursion is just partial recursion with the termination requirement added, and Turing [217] proved that there was no total computable function that could determine whether any given partial computable function call would terminate.

B.7 Enumeration and Indexing

Definition B.15. A "computably enumerable" set is a set for which there is a partial computable function whose value is defined if and only if its input argument is in the set, and undefined otherwise.

It is possible to enumerate the members of a computably enumerable set with an effective procedure, but it may not be possible to likewise enumerate the nonmembers.

Theorem B.16. *A set that is both computably enumerable, and whose complement (the set of nonmembers) is also computably enumerable, is computable.*

Proof. If a set is computably enumerable, there exists a partial computable function whose value is defined if and only if its input argument is in the set (thus verifying that members of the set are indeed members). If the set's complement is computably enumerable, then there is likewise a partial computable function whose value is defined if and only if its input argument is *not* in the set (thus verifying that *non-members* of the set are indeed not in the set). Therefore, if a set and its complement are both computably enumerable, we can combine the above two mentioned partial computable functions into a total computable function that returns 1 if its input is in the set, and 0 if it is not, since any given natural number must be either a member or a non-member of the set, and the value of exactly one of the two above-mentioned partial functions will always be defined; it follows, therefore, that the set is a computable set. □

Since the partial computable functions, and their calls, are defined in a purely constructive fashion (unlike the total computable functions) it is possible to assign a specific, computable index number to each one; so the partial computable functions (and their calls) are computably enumerable.

Definition B.17. An "alphabet" is any finite set of symbols.

Definition B.18. An "expression" is a finite combination of symbols taken from a given alphabet.

Definition B.19. A "language" will mean a set of "formulae", meaning allowed or "well-formed" expressions in some alphabet, or the set of the indices or Gödel numbers (see below) of these well-formed formulae.

Definition B.20. $\lceil A \rceil$ is a unique computable index or "Gödel number" for formula A, such that there is a computable function mapping from Gödel numbers to formulae, and an inverse computable function mapping from formulae to Gödel numbers.

Definition B.21. $\lfloor k \rfloor$ is the formula whose Gödel number is k.

Definition B.22. A "computably enumerable", "semi-recursive" or "Turing-recognizable" language is a language for which there is a partial computable function that returns 1 for input argument x if $\lfloor x \rfloor$ is a well-formed formula (but is not necessarily defined if $\lfloor x \rfloor$ is non-well-formed).

To specify a particular indexing, we attach a super-script, as in $\lfloor k \rfloor^M$ (for Gödel-numbering scheme M). (When the superscript is omitted, the particular indexing scheme is immaterial to the discussion at hand, or is assumed from context.)

So we might have, for instance, under a Gödel-numbering of partial functions,

$$k = \lceil F \rceil \tag{B.6}$$

$$F(x) = \lfloor k \rfloor (x) \tag{B.7}$$

or, under a Gödel-numbering of function calls,

$$k = \lceil F(x) \rceil \tag{B.8}$$

$$F(x) = \lfloor k \rfloor \tag{B.9}$$

A special partial computable function is the "universal Turing Machine", as defined in the main text, capable of simulating the behaviour of any other Turing machine (or partial computable function), given its index and a value for its input argument(s).

Partial computable functions of more than one argument can be reduced to single-argument functions, by using $\Diamond()$, a family of projection functions defined below. So $F(x_1, x_2)$ could be considered short-hand for $F(\Diamond(x_1, x_2))$. An even shorter form for the same thing will be $F(\langle x_1, x_2 \rangle)$. Likewise, $F(x_1, x_2, x_3) = F(< x_1, x_2, x_3 >)$, and so on. Strictly speaking, the n-place projection function will be called $\Diamond_n()$, but when the number of input arguments is clear, this subscript can generally be dropped.

So we will define a 2–D recursive projection function that maps from a pair of natural numbers (x_1,x_2) to a single natural number $\Diamond(x_1,x_2)$. Imagine that the number pairs are laid out as a 2–D matrix, where x_1 is the x_1^{th} row and x_2 is the x_2^{th} column, counting columns left-to-right and rows top-to-bottom. We then can convert the pair (x_1,x_2) into its number k by counting the elements of the finite (positive-slope) diagonals up to and including (x_1,x_2). There will be $(x_1 - 1) + (x_2 - 1)$, or $x_1 + x_2 - 2$, such diagonals that are complete. The sizes of the diagonal start at 1 and increase by one for each diagonal. In addition, we must add just x_2 for the final diagonal (rather than the $x_1 + x - 2$ we would count if it were a complete diagonal), since we are counting in this diagonal from the bottom left up and to the right just until we reach (x_1,x_2). Thus, we need to count the following number of elements, recalling that $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$:

$$\begin{aligned} \Diamond_2(x_1, x_2) = \sum i = 1x_1 + x_2 - 2i + x_2 & \\ &= \frac{(x_1 + x_2 - 2)(x_1 + x_2 - 1) + 2x_2}{2} \\ &= \frac{x_1^2 + x_2^2 + 2x_1x_2 - 3x_1 - x_2 + 2}{2} \end{aligned} \tag{B.10}$$

This provides a 2-place projection function only. Generalizing to n dimensions, we use the identity function for $\Diamond_1(x_1)$, and for $n = 3$, we use $\Diamond_2(x_1, x_2)$ as the basis to project two of the three dimensions onto one, and then another application of the same function gives us a 3-dimensional projection function, and so on for higher dimensions:

$$\Diamond_k(x_1, \ldots, x_k) = \begin{cases} x_1 \text{if } k = 1, \\ \frac{x_1^2 + x_2^2 + 2x_1x_2 - 3x_1 - x_2 + 2}{2} \text{if } k = 2, \\ \Diamond_2(\Diamond_{k-1}(x_1, \ldots, x_{k-1}), x_k) \text{ otherwise.} \end{cases} \tag{B.11}$$

I.e.,

$$\langle x_1, x_2, x_3, \ldots, x_n \rangle = \langle \ldots \langle \langle x_1, x_2 \rangle \rangle, x_3 \rangle, \ldots, x_n \rangle$$

B.8 The λ-Calculus

Much of the foundations of mathematics and logic becomes much easier when we adopt a foundational language in which functions can act as data, and any piece of data can be used as a function. In the λ-calculus we finally have an analytic language in which this is a built-in and natural feature. Compared to Turing machines, Church's formulation [54, 55, 56, 57] is more like predicate logic, as it is based on the idea of variable substitution. It is however, much simpler, with no reliance on propositional semantics or external reference. In the first real example of Turing-complete translation, Turing showed that Turing machines and the λ-calculus are formally equivalent [217, 218]. The λ-calculus is based on the idea of "abstraction", or filling in a form. Let us start with a specific example—let's say an application form for insurance, which can be filled in with many different combinations of names and addresses. When John Smith and Jane Doe fill in the *same* form, they end up with two different particular instances of it (which is why we call it abstraction). To represent John's filling in of the form, we can write:

$$Name: \_\_\_\_\_ \longrightarrow Name: \texttt{John Doe} \tag{B.12}$$

where "⟶" represents the "filling in", "evaluating" or "application" process.

A blank "_____" is an input "variable", or argument. A "form" is just a finite list (or ordered set) of elements contained within parentheses, such as $(a\,b\,c)$, known as a "λ-expression". The elements of the list can themselves be further lists, as in $(a\,b\,(c\,d)\,e)$, to any degree of complexity. There is also a standard way to indicate what pattern of characters to use for the "blank" or variable, which also allows us to have more than one variable. A λ-expression that can be evaluated takes the form

$$\lambda x.PI \tag{B.13}$$

where x, P and I are arbitrary λ-expressions; x being the variable, P being the expression to evaluate, and I being the value to bind to variable x. Application of $\lambda x.P$ to I results in the substitution of I for every instance of x in P:

$$\lambda(\_\_\_\_\_).\ (Name: \_\_\_\_\_)(\texttt{John Doe}) \longrightarrow (Name:\ \text{John Doe}) \tag{B.14}$$

A λ-expression can have any degree of nesting of applications within applications, and an application can generate a whole new application that was not there before. It is possible to construct expressions whose evaluation keeps "filling in the form" over and over again, recurring many times, or even looping forever. The following example loops forever by evaluating to a copy of itself:

$$(\lambda x.(xx))(\lambda x.(xx)) \longrightarrow (\lambda x.(xx))(\lambda x.(xx)) \longrightarrow (\lambda x.(xx))(\lambda x.(xx)) \longrightarrow \cdots \tag{B.15}$$

Any piece of a structure can be substituted with something smaller, even the empty list if necessary, as in

$$(\lambda y.xyz)() \longrightarrow xz \tag{B.16}$$

Although not always presented in functional terms, the λ-calculus can be thought of as being all about functions. Any $(\lambda x.P)$ is essentially a function with x as the argument, as in $P(x)$. $((\lambda x.P)I)$ is the function call or application. But a function call or application can also be itself a function, since we can apply it to yet another expression. There is no fundamental distinction here between functions, arguments and function applications; they are all just λ-expressions.

There are numerous ways to capture the operations of arithmetic and the natural numbers in the λ-calculus. Both Church and Kleene defined the naturals as follows: [56, 117]

$$\begin{aligned} 0 &\Leftrightarrow \lambda fx.x \\ 1 &\Leftrightarrow \lambda fx.fx \\ 2 &\Leftrightarrow \lambda fx.f(fx) \\ 3 &\Leftrightarrow \lambda fx.f(f(fx)) \\ &\vdots \end{aligned} \tag{B.17}$$

From here, one can program all the familiar devices of predicate logic, construct expressions that correspond to the Peano axioms of arithmetic, or the ZFC axioms of set theory, and derive and prove theorems: in short, do anything one can do in conventional mathematics or logic. One could do the same with Turing machines, of course, but the built-in structure of Church's calculus is more suited to a predicate logic interpretation.

An online λ-calculus interpreter can be found at [197] and a Lisp interpreter designed for foundational work can be found at [52].

B.9 Combinatory Logic

Perhaps the most elegant formulation of recursion is combinatory logic [120, 193, 59, 61, 62]. It also pre-dates all the other major languages (excepting the Analytical Engine). It is, in at least one way, even simpler than the λ-calculus, since it lacks the idea of variable bindings. A combinatory logic expression is a nested structure of ordered sets, just like a λ-calculus expression, but where the atomic symbols are also operators. There are two: S and K.

The S and K transformations are defined as follows:

$$\begin{aligned} Sxyz &\longrightarrow xz(yz) \\ Kxy &\longrightarrow x \end{aligned} \tag{B.18}$$

where x, y and z represent any sequence of bracketed S's and K's with balanced parentheses. So, for instance, KSS is transformed into S in a single step.

Given its lack of abstraction, lack of any function-data distinction, and extreme simplicity, I will take the SK-calculus as being very close to an ideal analytic basis language. The only possible reason to prefer the λ-calculus is that the SK-calculus does require *two* distinct operators. However, it is possible (as Shönfinkel himself recognized) to reduce these two further to a single operator, such as Barker's [13] iota operator ι:

$$\begin{aligned} \iota x &: xSK \\ f(x) &= x(g) \end{aligned} \tag{B.19}$$

Shönfinkel had his own single-operator version of combinatory logic, as well. However, he saw this reduction as "arbitrary", providing merely a syntactical simplification at the cost of semantical obfuscation.[72] Likewise,

[72] Quine [193, Intro.] questions this conclusion, asking where one is to draw the line—who is to say that *any* reduction in syntactical complexity is not "true simplification"? However, I will side with Shönfinkel on this one: it is always possible, in any language, to reduce syntactical complexity by obfuscating or over-complexifying the semantics. Shönfinkel's single-operator combinator, J, has a very contrived definition, and it is difficult to see how it is an improvement over S and K, with their very intuitive conceptualizations as "building-up" and "breaking-down". Similarly, Shönfinkel also shows how the left and right parentheses can be reduced to a single symbol, as well (and Quine points out that it is thus easy to produce a binary-coded combinatory language), but Shönfinkel does not see this as an improvement of any significance, either (does anyone really come away confused by the use of two structural punctuation marks, instead of one? Reducing this to a single punctuation mark still produces the same structures within structures that parentheses do, so where is the simplification?—the only real result seems to be that it is

the single operator of the λ-calculus is clearly a more complex operation than either S or K, since it requires abstraction/variables; however it is only one operator, as opposed to two, and is still quite easy to understand. Whichever system one prefers, the λ-calculus and the SK-calculus are very similar and closely related. Combinatory logic is, in fact, often referred to as "λ-calculus without abstraction", and the two systems are similar enough that much of the literature will mix the notations and terminologies together, as if they are simply different ways of talking about the same thing. The translation manual for going back and forth between the two is therefore relatively straightforward. From combinatory logic to the λ-calculus, we can translate as follows [60]

$$\begin{aligned} K &: \lambda xy.x \\ S &: \lambda xyz.xz(yz) \end{aligned} \tag{B.20}$$

And in the other direction:

$$\begin{aligned} \lambda x.y &: Ky \\ \lambda x.(ab) &: S(\lambda x.a)(\lambda x.b) \end{aligned} \tag{B.21}$$

The programming language JOY is based on combinatory logic, and the language UNLAMBDA is almost a direct implementation of the SK-calculus as a very simple programming language. (In addition, any programming that is called "functional" is generally modelled fairly directly on either the λ-calculus or combinatory logic.) An online combinatory logic interpreter can be found at [216].

B.10 Turing machines

Turing machines [218] (A. M. Turing, 1937) are the formalism often used when the mechanical, machine-like aspect of formal systems is being emphasized. Turing asks us to imagine idealized computing machines that read and write marks on a paper tape according to a set of internal instructions. The tape is a long one-dimensional string of little squares. The tape corresponds to a computer's memory and the list of instructions to a computer program. The tape must be, in theory, of infinite length (so your desktop PC is not *quite* a Turing machine, since it lacks an infinite memory).

The machine has a tape reader that points at one of the squares of the tape and can read the symbol that is written there. At each step in the execution of the program on the paper tape, the machine looks at the current square under its reader, and consults its list of instructions to see what to do. The machine can be in any of a finite number of different internal "states" (independent of whatever is written on the tape). The machine compares *(a)* the machine's current state, to *(b)* the current symbol it reads off the tape. Based on the result of the comparison, the program instructions tell the machine which of four possible actions to perform:

1. move the tape one square to the left,
2. move one square to the right,

much harder to visually discern the structure when examining an expression). Given that our purpose in developing a foundational language is to produce something with elegance and simplicity to our rational intuitions, we clearly need to distinguish between such artificial simplification versus true simplification. This does not mean, of course, that it is always easy to tell the difference, and I will not make any final judgement here as to whether Barker's ι is a true improvement over S and K. However, Barker's own material does not provide an explanation of ι's semantics directly, only defining it in terms of S and K, and so he does not really even try to make a case for true simplification. Here is a good rule of thumb: if the semantics of an operator cannot be explained in English, in a single sentence simple enough to be readily understood by listening to it (without any external aids), and that does not have the immediate appearance of a mere conjunction of two or more operators, then you may have over-complexified your semantics, and you might be better off expressing your operator in terms of a larger number of more readily understood operators.

3. change the symbol (including possibly to the same symbol, *i.e.,* do nothing), or
4. halt the computation.

In addition, the instructions must tell the machine what new internal state to go to (this information is required for the next iteration of the machine's running). For example, an entry in the machine's list of instructions might look something be:

```
IF (STATE 4) and (READING 0) THEN
  PRINT 1
  MOVE LEFT
  GO TO STATE 2
```

which, if the machine is in state #4 and the read head is seeing a "0" symbol, will print a "1" (overwriting the "0" it just read), move the tape one square to the left, and then put the machine into state #2.

One way in which Turing machines are useful is that they encourage a less conventionally "mathematical" intuition about recursive structures, helping us to view them more mechanically and without some of the synthetic artifact that we take too much for granted in conventional mathematical formulations. For instance, it is clear with Turing machines (unlike partial computable functions) that the *functional* aspect of recursion (the mapping of numbers onto numbers) is a purely interpretational gloss. When we view the structure as a machine, we can easily see that it is entirely arbitrary which squares we choose to isolate and "baptize" as the input arguments, and which ones we consider to be the "output" of the function. It is also perhaps intuitively easier to understand that the symbols on the tape are not inherently "numbers", but that this is merely a possible way of interpreting them.

While the mechanical flavour of Turing machines can be advantageous, they are still not the simplest formulation of recursion, and the concrete metaphor that they use is really just as much a synthetic artifact as the intuitions encouraged by more conventional mathematical notations. Also, there is still a distinction with Turing machines between *program* and *data*, which is unnecessary in a Turing-complete language.

B.11 First-order Predicate Logic plus Set Theory

Predicate logic with set theory[73] has become a kind of *de facto* standard in the foundations of mathematics. However, it has, in my opinion, far too much synthetic artifact to serve as a foundational analytic language, although it clearly has great usefulness as well as historical importance. Predicate logic is the most accepted general system for doing logic, serving (if anything does) as the "standard" logic. Set theory describes an ontology of "sets", which are widely seen as the primitive objects of mathematics. Thus, the two systems together (set theory implemented as a set of axioms in predicate logic) provide a good model of conventional mathematical reasoning. However, if our assumption of the analytic completeness of recursion is correct, there are many aspects of this system that impose arbitrary intuitional baggage on our analyses, making it inappropriate as an analytic basis language. These include: *(1)* the artificial distinction between logic and set theory, another example of the unnecessary function-data distinction, and *(2)* the use of predicates

[73] "Set theory" is added to predicate logic in the form of the set theoretic axioms. This usually means either Zermelo-Fraenkel (ZF) set theory or ZFC set theory (which is ZF set theory with the addition of the axiom of choice). The axiom of choice is usually assumed by those who wish to include completed infinities, such as the continuum, in their mathematics. While I have already stated that I consider completed infinities to be highly synthetic, this in no way means that I reject the use of ZFC set theory, since I consider both predicate logic and set theory to *already* be highly synthetic in nature, regardless of the adoption of the axiom of choice. In my opinion, neither is an appropriate *foundational* language for analysis: not for mathematics, logic or computation, all of which can be built up from the more solid analytic basis of combinatory logic. But this in no way makes predicate logic or set theory inappropriate fields of study, so long as their synthetic nature is understood. It is when ZFC set theory is held forth as *the* foundation for all mathematics—what mathematics is ultimately and objectively about—that I think predicate logic and set theory becomes enemies of rationality.

(functions that return either "true" or "false"), giving the system an inherently propositional semantics. Thus, predicate logic expressions are always propositions "about" something external to the system (there is nothing about a Turing machine, by contrast, that "refers" to anything outside of itself, not even to numbers, or to "true" and "false)". I would wager that the vast majority of those who see sets as the primitive objects of mathematics (probably a majority of mathematicians) either *(1)* see computation and logic as each *formally* distinct from mathematics (I would disagree with this position) or *(2)* accept that they are all formally equivalent but recognize the syntheticity of mathematics (while this is not the convention I have chosen to use here, I have no real beef with this usage).[74]

The programming language Prolog is based on first-order predicate logic. An online attempt to catalog and formalize (down to a ZFC foundation) the significant first-order proofs of ZFC mathematics can be found at [141].

B.12 Sequential Boolean Logic

This is the logic used by computer engineers who program directly in computer circuitry. A computer circuit is a network of logic "gates" that are hooked up to each other with wires. There are numerous kinds of gates, but the most basic is the **NAND** (or **NOT-AND**) gate:

With two binary (0 or 1) inputs and one binary output[75], a **NAND** gate outputs a 1 *unless* both of its inputs are 1. The **NAND** operation is more generally known as the "Sheffer stroke", and it can be used to create all the other basic Boolean functions of standard propositional logic, such as **AND**, **OR**, **NOT, IMPLICATION**, and so on.[76] Sequential Boolean logic can be a very handy language for foundational work. It is far simpler than predicate logic and its relatives. There is a single operator, with a very concrete way to visualize it (as an actual piece of circuitry), and we simply connect any number of these components together in any way we like, without restrictions. There *are* caveats to this apparent simplicity, however, the most important of which is that the language still distinguishes between function and data: the gates are functions and the wires carry data[77]. Thus, I tend favour our next two options over NAND gates, as they are also conceptually very simple and have no built-in function-data distinction.

A gate is an actual digital device, and has two inputs and one output. Since they can be hooked up in any pattern desired, they can be made to form arbitrarily nested structures that recur. There are sixteen (2^4) possible 2-place boolean operators, although if we require our system to have only one operator, this effectively limits us to eight possibilities, since each of the sixteen operators pairs with another of the

[74] The problem with doing otherwise is that "mathematics" is often presented as the general study of form or structure, in which case in must be more general than set theory, and equivalent to analysis. Likewise, "logic" is often presented as the general study of correct reasoning, in which case it must be more general than truth-valued logics, and also equivalent to analysis. To get further into the debate is beyond the scope of my current interests, which is why I have simply assumed an arbitrary convention. For more details about the connections between recursion theory and predicate logic, see any good introductory textbook on mathematical logic, such as [21] (some of this material can be found indirectly in Appendices B-C).

[75] Actually, it is logically identical to use either **NAND** or **NOR**, since they are essentially the identical operator, but with the role of 0 and 1 reversed. However, if there are no other gates, or logical operators, in your language, then reversing the role of 0 and 1 (or true and false) has no logical affect. So **NAND** and **NOR** are analytically identical (so long as there are no other operators in use).

[76] Sequential Boolean logic is, however, different from standard propositional logic, since the output of an operator is allowed to feed back into the input of the same operator (*i.e.*, operators can be recursive). Without the ability to recur, **NAND** gates become what is called "combinational Boolean logic", which is more or less the same thing as propositional logic. This can be confusing, especially since "combinational Boolean logic" is often simply called "combinational logic", or even "combinatorial logic", which can easily be confused with "combinatory logic". Combinatory logic, however, is Turing complete, while combinational logic is not.

[77] There are also timing issues that must be dealt with, so that it is well-defined when the outputs of operators will become available as inputs to other operators. Practical computer circuits usually work by synchronizing all the gates to the timing of a global clock.

sixteen that is functionally equivalent if used as the lone operator in the system. **AND**/**OR**, for instance, are equivalent as lone operators, as well as **NAND** and **NOR**. This is why it is convenient to refer to **NAND**/**NOR**, when it is the sole operator, by a single name: the "Sheffer stroke".

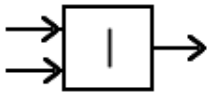
Of the eight remaining possible operators, the reader can readily verify that one more can be eliminated out of simple redundancy and three more because they completely ignore one of their two inputs, and so cannot serve as a fusion operator. Of the remaining, conjunction ($\wedge$), equality ($=$), implication ($\rightarrow$) and the Sheffer stroke ($|$), we want to keep only those operators that perform a *transformation* of their input data whenever a *match* occurs (*i.e.,* match-transform). An examination of the truth tables for these functions reveals that only the Sheffer stroke does this:

inputs	$\wedge$	$=$	$\rightarrow$	$\mid$
0 0	0	1	1	1
0 1	0	0	1	1
1 0	0	0	0	1
1 1	1	1	1	0

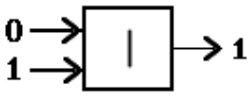
It seems easier, then, to justify the NAND gate in terms of the abstract "match-transform" notion, rather than the more concrete fusion operator of combinatory logic. Nonetheless, this is a relatively concrete way of thinking of abstraction, since the match operation is only done one bit at a time (although this one bit may be the result of the computation of a large and complex network of gates). However, it should also be noted that there is a sense in which, in spite of the concrete visual nature of the language, that sequential Boolean logic is still more abstract than combinatory logic. For instance, there is still a distinction here between function and data, so in that sense, it is not surprising that the match-transform concept seems more intuitively applicable. Hence, I will consider combinatory logic to be (probably) the purest analytic language we have, rather than sequential Boolean logic. However, NAND gates have their advantages too, including their close relationship with propositional logic, as well as their visual nature, which makes some kinds of discussions about pure analysis easier, especially for those who are visual thinkers.

The **NAND** gate is the most commonly used gate in actual computer circuitry, and is usually drawn as a combination of the **AND** gate with a tiny circle that represents the one-input **NOT** (or inversion) gate (although **AND** and **OR** gates are typically built out of **NAND** gates, rather than the other way around):

Since we are concerned here with foundations, and we only care to use a single operator, we do not have to distinguish between different kinds of gates, so I will draw our lone, universal logic gate as a box labelled with the Sheffer stroke symbol (a vertical line):

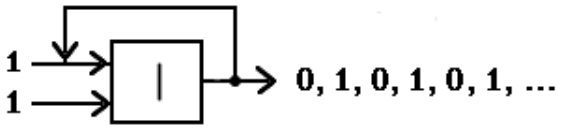


We could, of course, just use an empty box or circle, but I will keep the vertical line to remind us of what the gate actually does. I will label the input and output values as 1's and 0's, although I could just as well use *true* and *false* (we could also just as well switch around all our 1's for 0's and vice-versa). Since the operator performs a **NOT-AND**; an input of two 1's yields a 0 while anything else yields a 1:

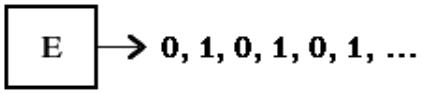


The boxes could also be thought of as functions, of course—in this case, functions of two arguments, but larger ensembles of **NAND**s can simulate functions of any number of arguments. This kind of functional diagram is, in general, called a "data-flow" diagram, since data flows along the lines, into and out of the functions represented by the boxes.

Not all **NAND**-based functions halt on all inputs, due to the presence of feedback. The following machine produces an infinite stream of alternating zeros and ones (and unlike the two earlier circuits, this one cannot be represented in combinatorial Boolean logic (logic gates without feedback ability), or propositional logic.

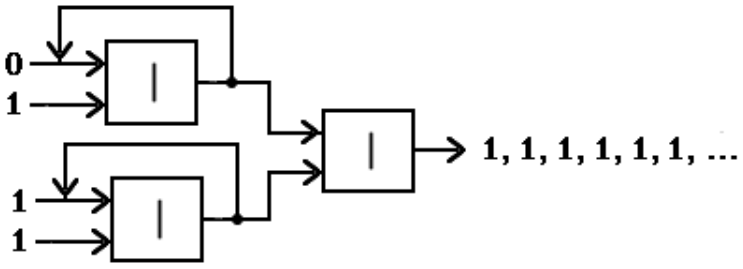


Let us call this the Epimenides function, draw a larger box around it, and label it thus:

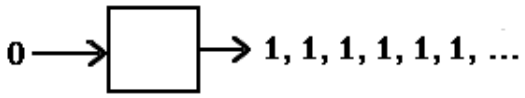


This grouping together of certain elements of a circuit, or program, into functional units like this, we will call "functional grouping". It has absolutely no effect on the computation that the circuit performs. It is there for our convenience, to make the whole thing more understandable. The same thing is done in the λ-calculus by giving certain expressions names and then using the names as a stand-in for the actual expression, to save on space. In the interpretation of a program as an "algorithm" or a "function" to solve some problem or other, or as a "proof" of something or other, we often need to perform this kind of functional grouping in order for our interpretation to make any kind of sense (for instance when we group together a sequence of squares on a Turing machine tape, and call it the "input"). Functional grouping, in any Turing-complete language, has no affect on the actual computation, so it is non-analytic or purely synthetic in nature.

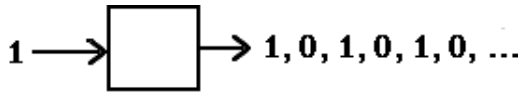
By combining operators into ever larger structures, we can program any recursive structure we like. The following function combines two Epimenides functions that cancel each other out, so to speak, producing an infinite stream of **1**'s:



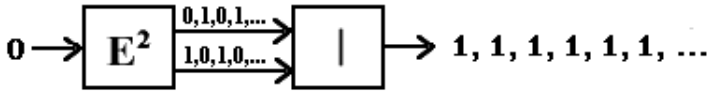
As another example of functional grouping, we might decide to call the above circuit a 1-argument function that produces all 1's when given a 0, as shown above:

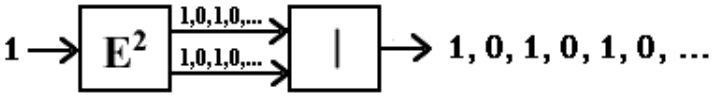


... but which produces alternating 0's and 1's when given a 1:



Or, we could chunk it down into two functions, one consisting of two Epimenides functions and having one input and two outputs—call it E^2—with the other being a single Sheffer stroke:





No matter how we choose to chunk our program up into "functions" and "output values", if the chunking has no effect on the computation that is performed, we will consider it to be synthetic interpretation, with no real analytic content. At bottom, even radically different functional groupings of the same circuit, which may seem superficially to "mean" very different things, are just the self-same conglomeration of many Sheffer strokes all hooked up to each other. This kind of synthetic "chunking" is exactly what is going on whenever we talk about "functions" and "algorithms", rather than just "programs". In fact, any talk about "functions" at all is essentially synthetic. Recall that some formulations of recursion—most notably, Turing machines—have no obvious interpretation in terms of functions at all. To view a Turing machine as a function or function call, one must rather artificially and arbitrarily "baptize" certain spots on the paper tape as "inputs" and others as "outputs".

This discussion of functional grouping applies equally to all the other Turing-complete languages, of course; I include the discussion under sequential Boolean logic only because its visual nature lends itself particularly well to such discussion.

B.13 Computer Programming Languages

This is actually a whole class of languages: those widely used for programming real-world computers. While these (mostly) lack the simplicity of more theoretical formulations, they foster the mechanical and constructive intuition that I am generally trying to encourage, while still being reasonably easy to read. Therefore, when actually presenting arguments about programs, I write my example programs in an informal pseudo-code language, `BASIC-F`, that is similar to popular procedural computer programming languages, like `BASIC`. All general-purpose computer programming languages are Turing-complete. Any programming language that was not Turing-complete would not be maximally expressive, and thus not much use as a general-purpose language.

A programming language, of course, is only technically Turing-complete if it has available to it an infinite memory store, something that is not the case for any actual computer—but this is a practical limitation on

physical computers, and not a limitation on the theoretical languages we invent for them (just as we do not have to worry about Turing machines having an infinitely long paper tape).

Dedekind's primitive recursion [65] is more or less equivalent to the use of `FOR` loops in modern programming languages, such as the one shown below in `BASIC-F`, which adds 1 to the variable `x`, 1000 times.

```
LET x = 0
FOR i=1 to 1000
    Let x = x + 1
NEXT i
```

Total and partial recursion is more or less equivalent to the use of `WHILE` loops, which contain a conditional halting test. For example, the Ackermann function [1, 156] could be coded as follows:

```
FUNCTION Ackermann(x, y)
  IF x=0 THEN
    RETURN y+1
  ELSE IF y=0 THEN
    RETURN Ackermann( x-1, 1 )
  ELSE
    RETURN Ackermann( x-1, Ackermann(x,y-1) )
  END IF
END FUNCTION
```

Unlike with a primitive recursive function, we cannot examine the above function and easily tell whether it will halt. The halting condition, or "base clause", is "`x = 0`", but it is not obvious that `x` will ever actually equal 0. In the case of the Ackermann function, it turns out that we actually can show that it will always halt, so it *is* computable. But one can quite easily create such functions that will not halt, or for which we just cannot tell whether they will halt. Ackermann's function was a very early example of a function that can be proved to be computable, and yet is *not* primitive recursive.

In fact, Turing's halting theorem [217] tells us that there is no way in general to distinguish between halting and nonhalting programs (*i.e.* between total and undefined partial recursive calls). This means that if we computably index (Gödel-number) all possible programs (*i.e.* create a code that allows us to give each program a unique number), there is no way we can write a function that will take the Gödel number of another function and return a value that indicates whether the Gödel-numbered function is computable or not (whether it will halt).

C Limit Recursion

C.1 Introduction

This appendix on limit recursion is intended to fill a number of gaps in the exposition of algorithmic information theory given in Ch. 4, in which the issues surrounding completed infinities, such as infinitesimals and the continuum, were largely swept under the rug. I gave general philosophical reasons for not considering the concerns about such infinities to be of very great substance—particularly since the dissertation is framed around a general assumption of the expressive adequacy of partial computable functions, which are inherently discrete. However, in recognition that there are many who feel differently, and consider infinitistic systems to be fundamentally more expressive than discrete systems, this Appendix lays out the most important reasons for believing discrete systems to be adequate. Limit recursion allows us to use Turing-complete languages to adequately express the truth (or falsity) of any statement that would require completed infinities to actually *prove* in a finite number of steps. This is accomplished by "proving" the statement "in the limit" of an infinite computation. This result does not invoke completed infinities, since the output of an infinite computation is never used as the input to another.

I will focus on the proof that limit recursion is adequate to do arithmetic, by showing that the arithmetic hierarchy (which contains all the statements of arithmetic number theory) is limit-computable (even though it is not computable). In the process of proving this, we will also see how first-order predicate logic can be subsumed under a Turing-complete language. This should clarify the relationship between traditional logic and computation systems (combinatory logic was, in fact, originally conceived as a foundation for predicate logic [193]).

Gold and Putnam [94, 165] first introduced the idea of limit recursion, which has been further developed by [86] and [33, 36]. It was introduced to account for some of the problems that come with taking either total and/or partial recursion as one's fundamental notion of an algorithm or effective procedure. There are problem-solving procedures that are of both practical and theoretical interest that do not fit easily into either the total or partial computable categories. A computable algorithm always halts and provides an answer. A nonterminating partial computable function call, on the other hand, does not halt, and hence never returns a value. The problem with this dichotomy is that a nonhalting procedure might well converge on some "output" value in the limit, if we were to monitor some part of its internal state while it was running. Thus, a nonhalting convergent process could be said to compute an answer "in the limit". This is not an answer that would be available to us to use (at least not with certain *knowledge* that we had the answer), but there might nonetheless be a sense in which the answer could be said to be "computed", and it is this possibility which we will now explore.

C.2 The Omega Rule

Putnam and Gold were not the first to use a notion of limit convergence, but were the first to develop it formally within the context of recursion theory. Similar ideas exist in proof theory that pre-date Putnam and Gold, namely certain systems of "infinite proofs", sometimes called "semi-formal systems".

Definition C.1. A "semi-formal proof" is any proof that makes use of the "Omega Rule", which assumes that, if there is an infinite sequence

$$F(0), F(1), F(2), \cdots \tag{C.1}$$

of sentences, all of which are provable, then we can conclude that

$$\forall x(F(x)) \tag{C.2}$$

The Omega Rule, or "Carnap's Rule", is not standard mathematical induction, although it appears superficially similar. Here, unlike in math induction, we must actually prove the infinite sequence of calls to $F()$ before we can draw our conclusion (something that it is obviously not possible to actually do, but with the Omega Rule, we presume that we can). Early hints of such systems pre-date Gödelian incompleteness [28, 234, 235]. After Gödel proved the incompleteness theorem, it was proved [107, 180] that the Omega Rule renders Peano Arithmetic complete:

Definition C.2. Theory "Z_Ω" is Peano Arithmetic (Z) with the addition of the Omega Rule as a valid rule of inference.

Theorem C.3. *Theory Z_Ω is complete (it has no undecidable sentences).*

This work was further developed in [2, 196, 219, 228, 136], and in a computational context by [10].

C.3 Limit recursion

Whereas the Omega Rule asks us to imagine that an infinite sequence of sentences can be used to infer a more general conclusion, limit-recursion asks us to imagine that an infinite sequence of computable function calls can be said to provide a more general result (in the limit). The Omega rule introduces a limiting device in he realm of proof theory, while "limit recursion" or "limit computation" does so in the realm of computation. With respect to the incompleteness theorem, the main significance of the former is to render Peano Arithmetic complete. The latter's significance is less clear, as it says nothing about what one can prove, but I will argue that constructing a limit-computation that corresponds to Gödel's theorem—which we will do later—can be a useful interpretational device.

In the limit recursion of Putnam, Gold and Burgin, a machine is constructed that computes an infinite sequence of $F(x)$ values, *i.e.,*

$$F(1), F(2), F(3), \ldots \tag{C.3}$$

This is possible, at least in principle, if $F()$ is computable and the machine has infinite time to run. If $F()$ is a predicate, then this infinite computation could be considered to correspond to the proof of "$\forall x F(x)$" by means of the Omega Rule. The conclusion, of course, never gets drawn here, so no proof has actually been effected by the running of this computation, but it does perform the infinite sequence of computations called for by the proof.

Gold calls the infinite sequence of values produced by this computation a sequence of "guesses", and states that "the problem will be said to be solved in the limit if, after some finite point in the sequence, all the guesses are correct and the same (in case there is more than one correct answer)." [94]. Putnam puts things in terms of computable sets, imagining a characteristic function of a set as returning a converging sequence of guesses as to whether a number is in the set, rather than returning a definite answer:

> "We know what sets are 'decidable', namely, the recursive [computable] sets (according to Church's Thesis). But what happens if we modify the notion of a decision procedure by (1) allowing the procedure to 'change its mind' any finite number of times (in terms of Turing Machines: we visualize the machine as being given an integer (or an n-tuple of integers) as input. The machine then 'prints out' a finite sequence of 'yesses' and 'nos'. The last 'yes' or 'no' is always to be the correct answer.); and (2) we give up the requirement that it be possible to tell (effectively) if the computation has terminated? I.e., if the machine has most recently printed 'yes' then we know that the integer put in as input must be in the set unless the machine is going to change its mind; but we have no procedure for telling whether the machine will change its mind or not." [165, p 49]

I will call this kind of recursion "limit recursion" or "limit computation", following [33, 37, 35, 36], but will generalize the term to a wider class of sets than just those covered by Putnam's term. Limit computable functions, like partial computable functions, will be a superset of the computable functions—every computable function can also be considered to be a limit-computable function that just happens to halt.

C.4 The Arithmetical Hierarchy

The arithmetical hierarchy can be an important tool in understanding limit recursion. It was created by [116] as a way of organizing arithmetic predicates into a hierarchy, with computable predicates at the lowest level, and each higher level being, in some sense "more uncomputable" than the last (we will see exactly what this means shortly). We will see in the next section that limit-computable functions can also be arranged in an exactly analogous hierarchy.

The arithmetic predicates in the arithmetical hierarchy are considered as "characteristic functions" defining sets of natural numbers.

Definition C.4. A single-argument natural number function $F(n)$ is a "characteristic function" of a set $\{n|F(n) = 1\}$.

Definition C.5. Any predicate $P()$ "defines a set" P such that for all natural numbers x,

$$x \in P \leftrightarrow P(x) \tag{C.4}$$

For convenience, we give the set the same name as the predicate that defines it, unless otherwise noted. We adopt the convenience of using natural numbers for both number values and truth values, so 0 is interpreted as false, while 1 (and any other value) is interpreted as true.

Definition C.6. The "decision problem" for set S is the problem of determining for any natural number n whether or not $n \in S$.

Definition C.7. The decision problem for a set is "solvable" or "computable" if its characteristic function is computable, otherwise its decision problem is "unsolvable" or"uncomputable".

The arithmetical hierarchy [116] allows for higher "degrees of uncomputability (or unsolvability)", and can be defined in terms of first-order predicate logic and computability, as follows.

Definition C.8. An expression of predicate calculus is said to be in "prenex normal form" if and only if it takes the form [116, p 50],

$$Q_1 x_1 Q_2 x_2 \ldots Q_n x_n E \tag{C.5}$$

where Q_k, for any natural number k, is a quantifier (either $\forall$ or $\exists$) with x_k being the single variable quantified over, and where n is the number of such quantifiers in the expression, and where E is an expression of first-order predicate logic with no quantifiers.

In other words, all the quantifiers are collected together at the very left of the expression without intervening parentheses. To put any arbitrary expression of predicate calculus in this form, all negation operators must first be placed inside quantified and parenthesized expressions. This can be done using the following standard logical equivalences:

$$\begin{aligned} \neg(A \,\&\, B) &= (\neg A \vee \neg B) \qquad & \neg(A \vee B) &= (\neg A \,\&\, \neg B) \\ \neg\forall x F(x) &= \exists x \neg F(x) \qquad & \neg\exists x F(x) &= \forall x \neg F(x) \\ \neg\neg P &= P \qquad & \neg(P) &= \neg P \end{aligned} \tag{C.6}$$

Once this is done, to put the expression in final prenex normal form, give each quantifier a unique variable name, to prevent naming conflicts, and then move all the quantifiers to the left of the expression, maintaining their order; what is now to their right is E in (C.5).

Any first-order predicate calculus expression can be put into prenex normal form, so it is never necessary to have quantifiers buried inside a predicate expression. We will restrict our predicate calculus to this syntax. If we refer to sentences that are not in prenex normal form, we shall assume that this is just a more informal way of referring to what are actually prenex normal form sentences.

Each level of the arithmetic hierarchy is a set of predicates, each defining a set of natural numbers. By using the Curry function $\Diamond()$, we could consider these predicates to be of any number of input arguments greater than zero, but it is customary to think of them as $2-place$ predicates of the form $P(k) = Q_1 x_1 \cdots Q_n x_n A(x1, \ldots, x_n, k)$, in prenex normal form—so $Q_1 \cdots Q_n$ is an alternating sequence of universal and existential quantifiers—and $A()$ is here interpreted as a recursive predicate (meaning, recall, a recursive function with output value 0 interpreted as false and all other output values interpreted as true) and k is any arbitrary natural number. Predicate $P()$ defines set S as follows:

$$P(k) = \begin{cases} 1 \text{ if } k \in S, \\ 0 \text{ otherwise.} \end{cases} \tag{C.7}$$

The number n is the "level" of $P()$ in the hierarchy, with $A()$ at level 0. Note that it follows from the above that for any predicate at level $n > 1$, a new predicate at level $n-1$ can be obtained by assigning an arbitrary natural number to x_1 and dropping $Q_1 x_1$.

The first level of the hierarchy (level 0) consists of all first-order expressions without quantification, with the unquantified $A()$ predicates interpreted as the computable predicates only. This means that *all* level 0 predicates are also computable, since there is clearly no way to create a partial computable expression out of first-order expressions employing only computable predicates.

Note that we are "interpreting $A()$ as computable", as this is a semantic property of our system, not an immediate feature of the syntax. The arithmetical hierarchy is a semantic interpretational tool, and so when we say anything about a predicate being computable or at such-and-such a level of the hierarchy, it should be understood that we a speaking within this particular metatheoretical context.

The sets at level 0 are (trivially) both computable and computably enumerable.

For convenience, level-0 (computable) sets (those defined by level 0 predicates) will be called by the same name as the predicates that define them, unless otherwise noted (so $P(x)$ defines set P).

Notice that, at level 1, we have all the predicate expressions which (in prenex normal form) have only one quantifier at the very left of the expression. At level 2, we have predicate expressions with two quantifiers, and then three quantifiers for level 3, and so on.

Definition C.9. A predicate is "universal" if the left-most quantifier in its the prenex normal form is a universal quantifier.

Definition C.10. A predicate is "existential" if the left-most quantifier in the prenex normal form is an existential quantifier.

While the level-0 sets are all computable, at level 1 we may have some sets which are computably enumerable but not computable. Consider the non-negated, universally quantified expression "$\forall x P(n,x)$", which defines the set $\{n|\forall x P(n,x)\}$.Speaking computationally, to use this universal predicate expression as a characteristic function to test if some number n is in the set, we would have to search through all natural numbers x, and check that $P(n,x)$ is true. If so, n is a member of the set. So it is clear that "$\forall x P(n,x)$" is not necessarily computable.

Recall that any universal quantifier can be re-expressed as a negated existential quantifier and vice-versa. So,

$$\begin{aligned} \forall x P(n,x) &\leftrightarrow \neg\exists x \neg P(n,x) \\ \neg\forall x P(n,x) &\leftrightarrow \exists x \neg P(n,x) \end{aligned} \tag{C.8}$$

Definition C.11. The set of existential level-1 predicates is "Σ_1".

Definition C.12. The set of universal level-1 predicates is "Π_1".

Since the negation (or complement) of a computable function is itself (trivially) also computable, we see that the set of level-1 universal functions (Σ_1) and the set of level-1 existential functions (Π_1) are complements of each other.

Theorem C.13. *Any predicate that can be expressed in both universal (negated-existential) and in existential (negated-universal) forms is computable.* [116][21, p.82]

Proof. Assume set A is defined by a universal predicate $A(n) = \forall x P_A(x,n)$ at level 1 of the arithmetic hierarchy (meaning that $P_A(x,n)$ is computable). Assume also that set B is defined by an existential predicate $B(n) = \exists x P_B(x,n)$ at level 1 of the arithmetical hierarchy (meaning that $P_B(x,n)$ is computable).

If an arbitrary natural number n is not in set A, we can find this out in finite steps by evaluating $P_A(x,n)$ for each possible value of x until we find a counter-example. If, however, n is in A, then we cannot evaluate $A(x,n)$ for every possible value of x, since there are an infinity of possible values of x, and we will never find a counter-example.

Likewise, if n is in set B, we can find this out in finite time, by evaluating the predicate for each possible value of x until we find a single case. If, however, n is not in B, then we cannot evaluate $P_B(x,n)$ for every possible value of x, since there are an infinity of possible values of x, and we will never find a valid case.

Assume that $A(x)$ and $B(x)$ are logically equivalent. Thus, $A(x)$ can be re-expressed in existential form and $B(x)$ can be re-expressed in universal form. We can now evaluate both $A(x,n)$ and $B(x,n)$ as described above, for each possible value of x in turn, alternating between $A(x,n)$ and $B(x,n)$, *e.g.*,

$$A(0,n), B(0,n), A(1,n), B(1,n), A(2,n), B(2,n), \ldots \tag{C.9}$$

until we find either a case of $A(x,n)$ that is false (in which case n is *not* in the set), or a case of $B(x,n)$ that is true (in which case the number n *is* in the set). □

Definition C.14. Those predicates at level 1 of the arithmetical hierarchy that have both universal and existential forms will be called "1-computable"; they are members of the intersection of Σ_1 and Π_1, which will be called "Δ_1".

By Theorem C.13, Δ_1 contains only computable predicates.

Likewise for the higher levels, we define Δ_i.

Definition C.15.

$$\Delta_i = \Sigma_i \bigcap \Pi_i \tag{C.10}$$

Theorem C.16. *The members of Δ_1 are exactly the computable predicates (so the 1-computable predicates are just the computable predicates).* [116, p 56]

For a set to be computable, recall that both it and its complement must be computably enumerable (Theorem 4.32). Therefore, a set whose defining predicate can be expressed only in universal/negated-existential form, being uncomputable, is not computably enumerable even though its complement is. Likewise, a set whose defining predicate can be expressed only in existential/negated-universal form is computably enumerable but its complement is not.

In fact, Kleene proved that

Theorem C.17. *Σ_1 is exactly the computably enumerable sets, and Π_1 is exactly the sets with computably enumerable complements.* [116, pp 57–8].

Theorem C.18. *Hierarchy Theorem: The Π_k and Σ_k sets, for all natural numbers k, form a strict hierarchy, so that each level k contains all the sets at the previous level $k-1$, plus additional sets. Furthermore, the Σ sets at each level contain sets not contained in the Π sets at the same level and vice-versa, and the Δ_k for an arbitrary level k contains all the sets at all lower levels in the hierarchy.*[116, p 49]

Those sets that are in Σ_k or Π_k, but not Δ_k, are all new sets for level k that do not appear anywhere in the lower levels $< k$.

Definition C.19. A set/predicate in Δ_k will be called "k-computable" or "k-recursive".

Definition C.20. A set that is in Σ_k will be called "k-computably enumerable" or "k-recursively enumerable".

So a set that is in Π_k has a k-computably enumerable complement.

To see why this last definition makes intuitive sense, note that while the Δ_1 sets are computable, the Δ_2 sets in general are not (they are 2-computable), but since they and their complements are computably enumerable relative to level 1, we will say that they are 1-computably enumerable. Likewise, the Δ_3 sets are 3-computable and 2-computably enumerable, and so on. The Hierarchy Theorem shows that well-defined, non-k-computable functions exist at all levels $k > 0$ of the hierarchy, ensuring us that this is not a vacuous exercise, but really is referring to new sets at each level that are not covered at lower levels.

C.5 Inductive Functions and Predicates

We now return to Putnam's and Gold's computation that makes an infinite series of "guesses". They develop this idea more formally using what they call "inductive" functions and predicates, which we will now define. My notation is mostly based on Putnam's. For convenience, wherever the variable x is used, the reader may read "$x_1, x_2, x_3, \ldots x_n$" for any $n \geq 1$.

Definition C.21. An "inductive function" or "partial inductive function" $f(x)$ is any function for which there is a computable function $r(x, y)$ so that, for all x, [165, pp 49-50]:

$$f(x) = k \quad \leftrightarrow \quad \lim_{y \to \infty} r(x, y) = k \tag{C.11}$$

$$\leftrightarrow \quad \exists y \forall z (z \geq y \rightarrow r(x, z) = k) \tag{C.12}$$

Definition C.22. A predicate $F(x)$ is an "inductive predicate" if, and only if, for some inductive function $f(x)$, we have $F(x) \leftrightarrow f(x) = 1$. Inductive function $f(x)$ is said to "define" inductive predicate $F(x)$, and is the characteristic function for the set defined by $F()$.

Definition C.23. When we say "by recursive induction on y", we mean "by the application of (C.11)".

Definition C.24. An inductive function $f(x)$ (and any inductive predicate $F()$ defined by it) is said to "converge" for a particular input x if its output value is defined for that x, *i.e.*, the limit operation in its definition ("$\lim_{y \to \infty} r(x, y) = k$" in (C.11)) has a determinate value for that value of x; otherwise $f(x)$ does not converge, and its output value is said to be "undefined" or "ϕ".

$$\lim_{y \to \infty} r(x, y) = \phi \quad \leftrightarrow \quad \neg \exists k \exists y \forall z (z \geq y \rightarrow r(x, z) = k) \tag{C.13}$$

Definition C.25. An inductive function $f(x)$ is a "total inductive function" and is said to be "convergent" if it always converges for all x. Likewise, an inductive predicate is a "total inductive predicate" if it is definable by a total inductive function.

In computational or procedural terms, it is useful to think of y informally as a kind of "counter" for an infinite sequence of repeated iterations of $r()$ evaluations or calls, performed over time (as we were informally conceiving of it earlier).

Although we have defined $r()$ as computable, following Putnam, it would be sufficient merely to require that it be undefined for no more than a finite number of y values, since that is all that is needed for it to converge.

Obviously, if a predicate $F()$ is computable, then the set it defines is also computable, and vice-versa.

Theorem C.26. *Set Δ_2 of the arithmetical hierarchy (the set of computably enumerable sets and their complements) is exactly the set of all sets definable by total inductive predicates.* [165, p 49]

Proof. Assume A is a computably enumerable set (and is thus in Σ_1). Now consider a function $g_A(y)$ that returns the y^{th} number in A, according to some enumeration scheme (the precise scheme chosen is immaterial). Function $g_A(y)$ is obviously computable. Function $f_A()$ is defined in terms of the graph of $g_A()$,

$$f_A(x, y) = \begin{cases} 1 & \text{if } x = g_A(y), \\ 0 & \text{otherwise.} \end{cases} \tag{C.14}$$

is also computable, and defines an inductive predicate $A(x)$ that tells us whether $x \in A$. Thus, any computably enumerable set can be defined as an inductive predicate.

We can now do likewise for B, rather than A, a set in Π_1, whose *complement* is computably enumerable. So the problem is in determining that certain numbers that are in B really are. Defining $g'_B(y)$ as the y^{th} natural number that is *not* in the set, the function $f_B()$ is defined in terms of the graph of $g'_B()$,

$$f_B(x,y) = \begin{cases} 0 & \text{if } x = g'_B(y), \\ 1 & \text{otherwise.} \end{cases} \tag{C.15}$$

and defines the inductive predicate $B(x)$, which tells us whether $x \in B$.

Thus, any set that is either in Σ_1 (being computably enumerable) or in Π_1 (having a computably enumerable complement) can be defined by an inductive predicate. Since $\triangle_2 = \Sigma_1 \cup \Pi_1$, it follows that any set that is in $\triangle_2$ can be defined by an inductive predicate.

To prove the converse—that any arbitrary total inductive predicate defines a set in $\triangle_2$–we start with an arbitrary inductive predicate $A(x)$ defined by some computable function $f(x,y)$. Since $f()$ is computable and returns only 1 or 0, it can be taken either as $f_A()$ in (C.14) or $f_B()$ in (C.15), defining inductive predicate $A()$ and its computably enumerable set A, or $B()$ and its set B with a computably enumerable complement. □

But what about the 3-computable sets, and the other k-computable sets for $k > 2$? How do these relate to the inductive predicates? Since the 3-computable predicates include predicates that are not 2-computable (by Theorem C.18), and since the inductive predicates are exactly the 2-computable ones, it follows that there are 3-computable functions that are not covered by inductive predicates. These are the ones that are 2-computably enumerable, while their complements are not (and, vice-versa, the sets that are not 2-computably enumerable but whose complements are). Although the predicates at this and higher levels are not all covered by the inductive predicates of Putnam and Gold, the idea of recursive induction has been generalized up the remaining levels of the arithmetical hierarchy [86, 33]. In outline, this basically involves replacing the computable function $r()$ in (C.11) with an inductive function, essentially applying recursive induction within what is already an application of recursive induction, in order to go up one level in the hierarchy, and then repeating the process to continue to higher levels. We will now look at this process in greater detail.

Definition C.27. When a recursive induction is performed on a variable y, as per Eqn C.11,

$$f(x) = k \leftrightarrow \lim_{y \to \infty} r^n(x,y) = k \tag{C.16}$$

except that the $(n-1)$-inductive function $r^n(x,y)$ replaces $r(x,y)$ (which we will now also refer to as $r^1(x,y)$), for $n > 1$, then $f(x)$ is an "n-inductive function" (where "1-inductive" just means inductive, and 0-inductive means computable).

Definition C.28. An n-inductive predicate, $F(x)$, is any predicate for which there is an n-inductive function $f(x)$ so that $F(x) \leftrightarrow f(x) = 1$ for all x, and where a "1-inductive predicate" is an inductive predicate and a "0-inductive predicate" is a computable predicate.

Definition C.29. An n-inductive function or predicate "converges" for a particular value of x if the function $r^n(x,y)$ in (C.16) converges for all values of y. The n-inductive function or predicate is "convergent" if it converges for all values of x.

The terms "total" and "partial" apply to n-inductive functions and predicates exactly as they do for the inductive ones.

We have already discussed how to compute, in the limit, the value of an inductive function (or predicate) call—namely, the machine computes an infinite sequence of calls to computable function $r()$, and this infinite sequence is interpreted as computing a value in the limit (a defined value if the sequence converges). This makes intuitive sense, since any given call to $r()$ in the sequence will eventually be computed, even though the sequence of calls will never be completed in a finite number of computational steps.

We can extend this notion of limit computation to k-inductive calls by first allowing that an infinite sequence of inductive calls can be interpreted as limit-computing the result of a 2-inductive call, just as we already accepted that a sequence of computable calls can be said to limit-compute the result of an inductive call. This makes intuitive sense only if our machine will eventually compute (as it did for the inductive functions/predicates) every $r^1()$ and $r^2()$ call specified by the definition of the 2-inductive function/predicate. Thus, our machine must compute an infinite sequence $r_1^2(), r_2^2(), r_3^2(), \ldots$ of infinite sequences $r_{k1}^1(), r_{k2}^1(), r_{k3}^1(), \ldots$ of computable calls, one infinite sequence for each $r_k^2()$. This forms an infinite 2-D matrix of computable calls, which can be traversed by indexing with $\diamondsuit_2()$, so the machine *will* eventually compute any given computable call in the resulting 2-D matrix of calls, even though it will never complete them all.

Precisely the same reasoning applies for n-inductive calls more generally, which will require our machine to traverse an n-D matrix of computable calls, which can be accomplished by indexing them with $\diamondsuit_n()$. This lends intuitive support to the extension of the idea of limit-computation of inductive calls to the n-inductive calls.

Note, however, that only in the case of inductive functions/predicates do we have, in general, the ability to compute an answer in a finite number of steps that thereafter never changes, and is correct. For an n-inductive call, the sequence $\ldots, r_k^n(), \ldots$ is the main sequence of guesses as to the output value being computed. In the convergent inductive case ($n = 1$), there is a value of k for which $r_k^1()$ returns the correct answer, which then never changes for higher values of k. But in the 2-inductive case (or higher), while each $r_k^2()$ call that we compute will eventually take on the correct value and never change, there will always be new such values we have yet to begin the computation of, and these will be subject to change before they converge. Thus, the overall computation will not, in general, produce the correct answer which then never changes—which remains true only for the 1-inductive functions/predicates.

Definition C.30. The evaluation of a "limit-expression", as on the right-hand side of (C.16),

$$\lim_{y \to \infty} r^n (x_1, \ldots, x_k, y) \tag{C.17}$$

will be called a "limit computation" if it converges, where "$x_1, \ldots, x_k$" is any number k of bound input variables, including possibly none, and where y is an additional optional bound input variable, and where $r^n()$ must have at least one input variable. (Note: this definition could be worded to suit any Turing-complete language.)

By convention, we will name limit-computations with capital Greek letters by writing the letter followed by a colon that the limit-expression:

$$\Lambda : \lim_{y \to \infty} r^n(x_1, \ldots, x_k, y) \tag{C.18}$$

Note that any n-inductive call can be interpreted as a limit computation, including calls to computable (1-computable) functions and predicates, since a limiting variable y does not have to appear in the argument list of $r^n()$.

Note that while "limit-computable" describes a function (or predicate or algorithm), a "limit-computation" is really just a computation—the uninterpreted mechanical process—that we are interpreting here in terms of limit-computability.

Theorem C.31. *Set $\triangle_3$ of the arithmetical hierarchy (the set of 2-computably enumerable sets and their complements) is exactly the set of all sets definable by total 2-inductive predicates.*

Proof. A Σ_1 set—which is computably enumerable—is 2-computable but not necessarily computable. Recall that we constructed its inductive characteristic function based on $f(x, y)$, which is 1 if $x = g(y)$ and 0 otherwise, where $g(y)$ is the yth number in the set. Now take a Σ_2 set, which is 2-computably enumerable; call it A_2. It is 3-computable but not necessarily 2-computable. Its characteristic function is defined by

$$f^2_{A_2}(x, y) = \begin{cases} 1 & \text{if } x = g^2_{A_2}(y), \\ 0 & \text{otherwise.} \end{cases} \tag{C.19}$$

where $g^2(y)$ is the yth number in the set, just as $g(y)$ was for set A (we drop the subscripts when they are not material)—except that this time, we are on level 2 of the arithmetical hierarchy, so the set is not computably enumerable, and thus $g^2()$ is not computable like $g()$ was, but is computably enumerable.

To generalize our (limit) computation of an inductive predicate (A) to one level higher in the arithmetical hierarchy (A_2), we can use our computation of the inductive predicate $A(x)$ to produce an infinite sequence of intermediate $f()$ values that converges on whether $x \in A$. Now *for each and every time* this $A(x)$ produces an intermediate $f()$ value, we take this result to be the final value of $A(x)$, as if A were computable. Assume we are on the kth $f()$ value. It follows from exactly the same reasoning as in our proof for the inductiveness of $\triangle_2$ predicates, that we can construct (for this intermediate value of $A(x)$) a computable $g_{A_2}()$ function, which in turn makes our $f_{A_2}()$ function computable, and our $A_2()$ predicate inductive. But these functions must now be computed for each different value of k, so they actually each form an infinite family of functions, which we will represent by adding k as an input argument to each, giving us the computable $g^2_{A_2}(y, k)$ and $f^2_{A_2}(x, y, k)$, and the inductive $A_2(x, k)$,

$$f^2_{A_2}(x, y, k) = \begin{cases} 1 & \text{if } x = g^2_{A_2}(y, k), \\ 0 & \text{otherwise.} \end{cases} \tag{C.20}$$

and where this defines the inductive predicate $A_2(x, k)$, which produces an infinite sequence of intermediate values (guesses) for the characteristic function of A_2, which converges on the right answer for whether $x \in A_2$ for some appropriately high value of k. This gives us two new 2-computable functions based on the above two computable functions, performing recursive induction on k. $F^2_{A_2}(x, y)$ will be the 2-computable function defined by $f^2_{A_2}(x, y, k)$, and $G^2_{A_2}(y)$ will be the 2-inductive function

$$\forall x \lim_{k\to\infty} f^2_{A_2}(x, y, k) = \begin{cases} 1 & \text{if } x = \lim\limits_{k\to\infty} g^2_{A_2}(y, k), \\ 0 & \text{otherwise.} \end{cases} \tag{C.21}$$

In other words,

$$\forall x\, F^2_{A_2}(x, y) = \begin{cases} 1 & \text{if } x = G^2_{A_2}(y), \\ 0 & \text{otherwise.} \end{cases} \tag{C.22}$$

So for each value of k in this induction, we get a separate, independent infinite sequence of output values for $f^2_{A_2}()$ and its corresponding 2-inductive predicate $F^2_{A_2}(x)$.

We can now do exactly the same thing with any Π_1 function, call it B_2 that has a 2-computably enumerable complement. It is, like A_2, 3-computable but not 2-computable, its 2-inductive characteristic function being defined by

$$F^2_{B_2}(x,y) = \begin{cases} 0 & \text{if } x = G^{2\prime}_{B_2}(y), \\ 1 & \text{otherwise.} \end{cases} \tag{C.23}$$

By recursive induction on k,

$$f^2_{B_2}(x,y,k) = \begin{cases} 0 & \text{if } x = g^{2\prime}_{B_2}(y,k), \\ 1 & \text{otherwise.} \end{cases} \tag{C.24}$$

This defines the 2-inductive, 3-computable $B_2(x)$ predicate, by an argument exactly analogous to the one for A_2. Thus, any set that is either in Σ_2 (being 2-computably enumerable) or in Π_2 (having a 2-computably enumerable complement) can be defined by a 2-inductive predicate. Since $\triangle_3 = \Sigma_2 \cup \Pi_2$, it follows that any set that is in $\triangle_3$ can be defined by a 2-inductive predicate.

To prove the converse—that any arbitrary 2-inductive predicate defines a set in $\triangle_3$—we start with an arbitrary 2-inductive predicate $P_2(x)$ defined by inductive function $g^2()$. Since $g^2()$ is inductive, it can be taken either as $g^3_{A_2}()$ in (C.21) or $g^{2\prime}_{B_2}()$ in (C.24), defining inductive predicates $A_2()$ and its 2-computably enumerable set A_2, or $B_2()$ and its set with a 2-computably enumerable complement. □

To summarize intuitively what we have accomplished thus far, we have now defined a way to compute "in the limit" all the arithmetical predicates up to and including level 2, by showing how to define them as 1- and 2-inductive predicates, using limit recursion. We generalized from 1- to 2-inductive by performing another recursive induction on a second variable, once for each "guess" made in the computation of the first induction, this being computable using the $\diamondsuit_2()$ function. We will now likewise generalize to show that

Theorem C.32. *Set $\triangle_{n+1}$ of the arithmetical hierarchy (the set of n-computably enumerable sets and their complements) is exactly the set of all sets definable by total n-inductive predicates.*

Proof. Define the n-inductive predicates corresponding to the $(n+1)$-computable predicates in the arithmetic hierarchy, for the Σ_n and for the Π_n sets, as we did before for $n = 1$ and $n = 2$.

$$F^n_\Sigma(x,y) = \begin{cases} 1 & \text{if } x = G^n_\Sigma(y), \\ 0 & \text{otherwise.} \end{cases} \tag{C.25}$$

By recursive induction,

$$\lim_{k_1 \to \infty \cdots k_n \to \infty} f^n_\Sigma(x,y,k_1,\ldots,k_n) = \begin{cases} 1 & \text{if } x = \lim_{k_1 \to \infty \cdots k_n \to \infty} g^n_\Sigma(y,k_1,\ldots,k_n), \\ 0 & \text{otherwise.} \end{cases} \tag{C.26}$$

Likewise,

$$F^n_\Sigma(x,y) = \begin{cases} 0 & \text{if } x = G'^n_\Sigma(y), \\ 1 & \text{otherwise.} \end{cases}$$

$$F^n_\Pi(x,y) = \begin{cases} 1 & \text{if } x = G^n_\Pi(y), \\ 0 & \text{otherwise.} \end{cases}$$

$$F^n_\Pi(x,y) = \begin{cases} 0 & \text{if } x = G'^n_\Pi(y), \\ 1 & \text{otherwise.} \end{cases} \tag{C.27}$$

Thus we have shown that the $(n+1)$-computable predicates/sets in the arithmetical hierarchy are all n-inductive.

To prove the converse—that any arbitrary total n-inductive predicate defines a set in $\triangle_{n+1}$—we reason just as we did for $n = 1$ and $n = 2$, starting with an arbitrary total n-inductive predicate $P_n(x)$ defined by total $(n-1)$-inductive function $g^n()$. Since $g^n()$ is $(n-1)$-inductive and total, it can be taken either as $g^n_\Sigma()$ or $g^{n\prime}_\Pi()$, defining inductive predicates $A_n()$ and its n-computably enumerable set A_n, or $B_n()$ and its set with an n-computably enumerable complement. □

So far, we have used the term "computation in the limit" informally, but we are now in a position to define it in terms of limit recursion (n-induction).

Definition C.33. The k-inductive functions, or predicates, that are convergent (total) will be called "computable in the limit" or "limit-computable" or "total limit-recursive". These, in addition to those that are not convergent (do not converge for all values of their input variables), will be called "partial limit-computable" or "partial limit-recursive".

Theorem C.34. *The predicates of the arithmetical hierarchy are all limit-computable (total k-computable for some value of k).* [34]

Proof. Immediate, from Definition C.33 and Theorem C.32. □

Note that our definition of "limit-computable" is broader than that of Gold's "limiting recursive" and Putnam's "trial and error predicates", both of which correspond only to what we are calling "inductive predicates", which applies only to Δ_2.

The way we have described the arithmetical hierarchy thus far, level 0 consists of the computable predicates. This would not, however, allow us to enumerate the AH predicates, since the computable functions are not enumerable. However, it can be shown that limiting level 0 to the primitive recursive functions is adequate to generate a hierarchy with exactly the same predicates [116, p 57]. In fact, we can restrict ourselves even further to the addition and multiplication functions and still end up with the same predicates. This follows from the fact that, in conjunction with predicate logic, addition and multiplication on the natural numbers are adequate to produce all computable functions, and are themselves computable [93, pp 34-5].

C.6 Pseudo-code Examples

The following BASIC-F code illustrates the use of limit computation. Note that these procedures are not inherently limit computations; they are just programs like any other. The idea that they are performing

limit computations is an interpretation of them as algorithms, not an inherent feature of them as programs. For more detail and further explanation of this code, see [176].

C.6.1 Gödel's Incompleteness Theorem

The following BASIC-F procedure limit-computes the truth of Gödel's incompleteness theorem: that there are truths of number theory (and similarly for any system at least as powerful) that are unprovable within number theory (assuming number theory's consistency). The non-limiting proof of this, given by Gödel [93], can be considered (as mentioned in the main text) to be an *a priori* analytic proof that there exist *a posteriori* analytic statements. The limit-proof demonstrates why the mere existence of such truths does not demonstrate the reality of a non-analytic ontology, as was more or less the (unpublished) view of Gödel himself [92]. This is of significance in this dissertation, because I have assumed the adequacy of analytic (specifically, computational) languages to cover any *a priori* ontology, in spite of the usefulness I have given to synthetic *a priori* epistemological methods.

```
PROCEDURE G // limit-computes the truth of Godel's incompleteness theorem
    FOR k=0 TO INFINITY
        PRINT 1
        IF U(k)=0 THEN
            ERASE(1)
            PRINT 0
        END IF
    NEXT FOR
END PROCEDURE

FUNCTION U(x)
    FOR i=0 TO INFINITY
        IF C(x,i)=1 THEN
            ERASE(i)
            RETURN 0
        ELSE
            PRINT 1
        END IF
    NEXT FOR
END FUNCTION

FUNCTION C(x,i)
    IF Z(x)=T(i) OR ~Z(x)=T(i) THEN
        RETURN 1
    ELSE
        RETURN 0
    END IF
END FUNCTION

FUNCTION Z(k)
    RETURN the Godel number of the kth Peano sentence
END FUNCTION

FUNCTION ~Z(k)
    RETURN the Godel number of the negation of the kth Peano sentence
END FUNCTION
```

```
FUNCTION T(k)
  RETURN the Godel number of the kth Peano theorem
END FUNCTION
```

C.6.2 The Arithmetical Hierarchy

The following BASIC-F function limit-computes an arbitrary arithmetical sentence, which will be written as having the form $QxE[x]$, where "Q" and "x" could both be null. Note that "$E[x]$" is not a function or call, but just represents a formula with instances of unbound variable "x" within it. When we write "$E[i/x]$", it means to substitute all instances of "x" in "$E[x]$" with "i".

```
FUNCTION LIMIT–COMPUTE(QxE[x])
   IF Q is existential THEN
      FOR i = 0 TO INFINITY
          IF LIMIT–COMPUTE(E[i]) = 0 THEN
             TRY 0
          ELSE
             RETURN 1
          END IF
      NEXT i
    ELSE IF (Q is universal) THEN
      FOR i = 0 TO INFINITY
          IF LIMIT–COMPUTE(E[i]) = 0 THEN
             RETURN 0
          ELSE
             TRY 1
          END IF
      NEXT i
    ELSE // there is no quantifier
      IF E = 0 THEN
         RETURN 0
      ELSE
         RETURN 1
      END IF
   END IF
END FUNCTION
```

D The BASIC-F Language

D.1 Introduction

This appendix provides a more complete description of the pseudo-code language used in the toy examples called BASIC-F, for "Beginner's Analytic Symbolic Instruction Code for Foundations." This language is obviously Turing-complete, and bears some similarity to well-known variants of the BASIC programming language. Despite its prominent use in the main text, I did not fully describe it there, as I felt it was self-explanatory enough that this would only distract the reader. However, I include a more detailed description here, for the sake of completeness, and for those unfamiliar with the conventions of standard procedural programming languages.

This is the same language I have used elsewhere to discuss the philosophy of mathematics [176]. Not all features of the language are used in the main text, although much of the remainder makes an appearance in Appendices B-C. Note that BASIC-F, while being in most respects a fairly standard pseudo-code language, has some features that make it particularly suited to foundational studies (such as `TRY`, `ERASE` and the dot "." notation).

Some may find these special "foundations" features to be perplexing in a dialect of BASIC, since it is generally accepted that languages like the λ-calculus and combinatory logic are more appropriate for foundational studies (at least for those that take an algorithmic approach). For instance, Chaitin uses a version of Lisp [51], which is closely based on the λ-calculus. However, I have found, that for my purposes, using such esoteric languages—while more appropriate from a purely theoretical perspective—does not necessarily yield a clear benefit, while at times bogging us down in the excruciating minutiae of implementing even simple computations in such stripped-down languages. For some work, it may be beneficial, but I would resist going that route unless there was a real purpose to it. Chaitin's Lisp is an admirable attempt to find a balance, by adding practical features above and beyond the pure λ-calculus, while at the same time keeping things within a stone's throw of its theoretical purity. However, the resulting programs are still difficult for a non-programmer to understand, and for my purposes, what matters most is that we understand that our programs *can* be translated into more analytically pure languages, and that we understand those purer languages well enough not to ascribe undue significance to synthetical artifacts of whatever language we actually decide to use.

I have still, of course, tried to avoid unnecessary synthetic artifacts that do not aid in understanding, and would take it as given that we wish to avoid any language that tends to come with an implicit external (non-constructive) or propositional semantics, such as predicate logic or set theory. From a theoretical perspective, it would also be nice to avoid all languages that employ a function-data distinction, but I don't think this is possible if we wish to avoid overly esoteric code.

I would, on the other hand, not hesitate to employ a purer language as soon as a particular need for it

arose, and I would encourage readers to view my pseudo-code programs as being merely a handy short-hand way of specifying SK combinators. For the sake of completeness, I have specified functions in BASIC-F to perform Gödel-numbering and to evaluate λ and SK expressions, but they are not used in this dissertation.

I have kept the specification of the language less formal than it could have been, to aid in understanding, but I believe it to be as formal as it reasonably needs to be. It is, I believe, specified formally enough to permit the actual implementation of a BASIC-F interpreter with a minimal amount of additional formality. It is, however, primarily intended for writing pseudo-code, so we also must allow for a certain amount of natural language description (*e.g.,* English), ambiguity, and even outright breaking of the rules, so long as no unnecessary ambiguity results. The whole point of a pseudo-code language is to *not* be shackled by excessive formality, while retaining enough formality to communicate the algorithm formally enough for the given purposes.

Note that, while this appendix expresses pre-defined BASIC-F keywords in all-capitals, no restriction to uppercase is intended in actual use.

`BASIC-F` contains no elements that are proprietary or particularly new, so you may feel free to use it, without attribution, to communicate your own algorithms.

D.2 Dictionary

I follow a number of fairly standard conventions here. For instance:

$$a \,|\, b \,|\, c$$

means that one of either a or b or c may appear here (and, of course, the convention works with any number of alternatives, not just three).

Any text inclosed in angle-brackets, such as

$$< label >$$

is intended to label or describe the actual text, rather being taken as verbatim text. Such a label is nothing more than a convenient label arbitrary text, unless specifically defined as having some special characteristics and constraints.

Optional text is enclosed in square brackets, so that

$$[a]$$

means that a may or may not actually appear.

Ellipses can be used to indicate arbitrary code, or to indicate that an obvious pattern repeats, as in

$$1, 2, 3, \ldots$$

although obviously there are no infinite patterns in a programming language, so the above would not be allowed for that reason. However, we *could* find

$$1[, 2[, 3[, \ldots]]]$$

which would be the same pattern, but continued to any optional finite length.

Note that the above conventions are not part of the language. They are merely conventions used in this sections for describing and defining the language.

Commands

"Commands" are those defined here as built into the language, *or* new commands defined in code using `PROCEDURE` (see below). Commands occur line-by-line, so that a carriage return signals the beginning of a new command.

```
<command1>
<command2>
...
```

A back-slash can be used, however, before the carriage return to indicate that the next line simply continues the current line. So we can have

```
<command1>
<command2 line1> \
    <command2 line2> \
    ...
<command3>
...
```

Multi-line commands using back-slashes exist only to allow readable formatting of very long commands. The back-slash gives the language no extra features or abilities.

Values

The special labels

$$< value >, < value1 >, < value2 >, \ldots$$

will always refer to one of:

1. a number (natural, integer, finite-precision decimal or complex numbers are allowed).
2. a block of code that evaluates to a number using standard arithmetical rules, including but not limited to:

$$\begin{array}{rcl} (< value >) & & \\ < boolean > & & \\ < value1 > & + & < value2 > \\ < value1 > & - & < value2 > \\ < value1 > & * & < value2 > \\ < value1 > & / & < value2 > \end{array}$$

3. a string, meaning any arbitrary text, enclosed in quotes when appearing in code, as in `"<value>"`, but appearing in output (from the PRINT command) without the quotation marks. If a string is used in an exclusively arithmetical context, its value will be taken to be its numerical value if it has a form that can be so evaluated, and zero otherwise (so the value of `"0+2"` or of `"DOG"+2` will be 2, but `PRINT "DOG"` will output `DOG`, not `0`).
4. an invocation of (or call to) a `FUNCTION` (see below), which must return a `<value>` as defined by one of the above.

Booleans

The special labels

$$< boolean >, < boolean1 >, < boolean2 >, \ldots$$

will always refer to one of:

1. `TRUE` or `FALSE`, which evaluate to true and false, respectively. If used in an arithmetical context, `TRUE` will evaluate to 1 and `FALSE` to 0.
2. Any $< value >$ where 0 is considered false and any nonzero is considered true.
3. a block of code that evaluates to true or false in the standard manner, including but not limited to:

$< value >$		
TRUE		
FALSE		
$< value1 >$	=	$< value2 >$
$< value1 >$	<	$< value2 >$
$< value1 >$	>	$< value2 >$
$< value1 >$	<=	$< value2 >$
$< value1 >$	>=	$< value2 >$
$< value1 >$	AND	$< value2 >$
$< value1 >$	OR	$< value2 >$
$< value1 >$	IMPLIES	$< value2 >$
$< value1 >$	IF	$< value2 >$
$< value1 >$	IFF	$< value2 >$
$< value1 >$	XOR	$< value2 >$
$< value1 >$	NAND	$< value2 >$
$< value1 >$	NOR	$< value2 >$
NOT	<value1>	

Variables

The special labels

$$< var >, < var1 >, < var2 >, \ldots$$

will refer to variables. A variable name can actually be any word not pre-defined by the language. I will also use $< arg >$, etc. as labels for variables that are input arguments to a procedure or function. Variables are permanent storage locations for values, and can be invoked anywhere that a `<value>` can appear.

Note that a variable's scope is only within the `PROCEDURE` or `FUNCTION` in which it is defined (where definition occurs by way of first use, with default values of zero). Thus if procedure `A` puts the number 3 in variable `X`, and then calls procedure `B`, which immediately uses variable `X`, then variable `X` in procedure `B` will have value 0, *not* a value of 3, since it is a different variable than the `X` that appeared in procedure `A`. If a procedure needs to access the variables of another procedure, the values must be passed as arguments by the calling procedure to the called procedure.

In addition, it should be noted that when a particular procedure or function is called, its variable space starts with a clean slate. It does not use the same space that previous calls to that procedure used. Hence, a function can even call itself, for instance, and this second invocation of the function will start with a clean slate.

Comments

```
...
 /* <some comments
    to explain the code
    on more than one line.>
 */
... // <a comment appended to a single line.>
...
```

The double-slash can be used to introduce a comment anywhere on a line, so that the remainder of the line is ignored when the program runs, having no functional consequences. The slash-asterisk opens a multi-line comment, which is closed by the asterisk-slash. Comments are usually used to aid the reader in understanding the actual code.

LET

```
LET <var> = <value>
```

Variables can be assigned values with the LET command. The above assigns the variable `<var>` the value `<value>` (remember that the labels in the angle brackets are simply descriptions of the actual text; there is no actual use of angle brackets in BASIC-F). An example of an actual LET statement might be:

```
LET X = 6.28
```

PROCEDURE

```
PROCEDURE <procedure-name> [ ( <arg1>[,<arg2>[,<arg3>[,...]]] ) ]
    ...
END PROCEDURE
```

The above defines a "procedure", which can contain any desired code, and can accept any finite number of input arguments (arg1, arg2, etc.), including possible none. A procedure allows ease of code re-use, and once defined becomes a command in the language, and can be called simply by invoking its name on a line, as if it were a built-in command of the language.

Note that the arguments are "call-by-value", meaning that <arg1> starts out with the value of whatever was passed to it from the calling procedure, but henceforth, within this new procedure, it is a different variable. If it is assigned a new value, this will not affect the original variable in the original calling procedure.

FUNCTION

```
FUNCTION <function-name> [ ( <arg1>[,<arg2>[,<arg3>[,...]]] ) ]
    ...
    RETURN <value> | TRY <value>
    ...
END FUNCTION
```

A function is just like a procedure (in fact, we will refer to it as a kind of procedure), except that it "returns" a `<value>`. Thus, when a function is called, its invocation is replaced by the value of the computation performed by the function, meaning whatever `<value>` is returned by either the `RETURN` or `TRY` commands within the function (a function's code *must* contain at least one `RETURN` or `TRY`, but may contain any number of them).

Note that as soon as a function hits a RETURN or `TRY` command, it halts and returns the value.

The `RETURN` command simply returns the given value back to the function's invocation in the calling procedure. The `TRY` command works the same as `RETURN`, except that if it occurs inside a loop (see `WHILE`, `REPEAT` and `FOR`), it only tentatively returns a value, while the execution of the loop continues in parallel with the returning of the `TRY` value. Each such execution is called a "thread". Thus, there can be, potentially, as many `TRY` values tentatively returned as there are loop iterations, all running as threads in parallel (although, of course, some iterations of the loop may actually `RETURN` values). Once a `RETURN` for a function call is executed, all previous `TRY` values for that loop are deleted, along with any further resulting processing.

(The `TRY` command is intended mostly for the implementation of limit-computations.)

PROGRAM

```
[ PROGRAM [ <program-name> ] ]
    ...
[ END PROGRAM ]
```

The above defines a "program", which is just a procedure with no input arguments. The program may be named `<program-name>`, but this is not necessary if the program is not going to be called by another program. Note that, since the input arguments to a procedure are entirely optional, the `PROGRAM` command is fully equivalent to using `PROCEDURE` without any input arguments. One cannot, however, use `PROGRAM` *with* input arguments, and one also cannot define an unnamed `PROCEDURE`, as one can with `PROGRAM`.

Any BASIC-F code that appears stand-alone, not inside a PROGRAM, PROCEDURE or FUNCTION definition, is interpreted as an unnamed program.

WHILE

```
WHILE <boolean>
    ...
END WHILE
```

The block of code represented by the "..." will loop (execute over and over again) so long as `<boolean>` is true. Its truth is checked at the top of each loop iteration (before the execution of that iteration). As soon as `<boolean>` is found to be false, execution immediately proceeds to whatever is immediately following the loop's `END WHILE`.

REPEAT

```
REPEAT
    ...
UNTIL <boolean> | WHILE <boolean>
```

The block of code represented by the "..." will loop either until `<boolean>` is true, or while it is true, depending on whether `UNTIL` or `WHILE` is used. The `<boolean>` is checked at the bottom of each loop iteration (after the execution of that iteration, and before the next iteration). The block of code will hence necessarily execute at least once. As soon as `<boolean>` is found to be true (for `UNTIL`) or false (for `WHILE`), execution immediately proceeds to whatever is immediately following the `UNTIL` or `WHILE`.

The above code is fully equivalent to the following `WHILE` loop:

```
<var> = TRUE
WHILE <var>
    ...
    LET <var> = [NOT] <boolean>
END WHILE
```

where the `NOT` is included if we want to use `REPEAT-UNTIL`, and omitted for `REPEAT-WHILE`.

FOR

```
FOR <var> = <var1> TO <var2> [ STEP <var3> ]
    ...
NEXT <var>
```

This is a loop where a variable `<var>` starts with a value of `<value>` and is automatically incremented by `<var3>` (which will simply be 1 if not specified) before each repetition of the loop. The above is fully equivalent to the following `WHILE` loop:

```
<var3> = 1
[ LET <var3> = <value> ]
<var> = <var1>
WHILE <var> <= <var2>
    ...
    LET <var> = <var> + <var3>
END WHILE
```

PRINT

```
PRINT <value>
```

The above evaluates <value> and outputs the result. One can imagine this output going to a printer or display screen.

ERASE

```
ERASE [ (<k>) ]
```

The `ERASE` command has the ability to erase things that have already been output by a `PRINT` command. If there is *no* input argument, then `ERASE` will simply erase the last output of a `PRINT` command. If there *is*

an input argument (`<k>`) and the `ERASE` occurs inside a thread created by a TRY command, then the results of all `PRINT` statements in the `<k>`*th* thread will be erased (meaning in the thread corresponding to the `<k>`*th* iteration of the loop that generated the thread). Thus, if one of the iterations did *not* use TRY (but simply returned a value with `RETURN` or did not return at all) then it is *still counted* in the thread enumeration, although its `PRINT` results cannot be erased with `ERASE` <k>, since it is not actually a true thread. More generally, any attempt to use `ERASE` <k> to erase the output of a non-existent thread will simply have no effect.

An `ERASE` can have effects beyond the scope of the immediate procedure it appears in. If it is running in a thread that is itself running inside another thread, it is the former (the inner-most) thread that will be the one the `ERASE` is applied to.

INDEX

... INDEX (<expression>, [<enumeration>]) ...

This functions returns the numerical index (Gödel number) of `<expression>` in an `<enumeration>`. The `<enumeration>` can be left out if some standard enumeration in some standard language is understood.

EXPRESSION

... EXPRESSION (<index>, [<enumeration>]) ...

This inverse-Gödel function returns the `<index>`*th* expression in an `<enumeration>`. The `<enumeration>` can be left out if some standard enumeration in some standard language is understood.

EVALUATE

... EVALUATE (<expression> [, <k>]) ...

This function takes the `<expression>` string, which it interprets as a BASIC-F, SK-calculus, or λ-calculus program (whichever is possible, in that order of preference), and returns the value or output (as a text string) from `<k>` program steps (where a program step is one application, or the execution of one line of code for BASIC-F). If no `<k>` is specified, it returns the final output after application has halted (if the expression does not halt, then the call will never halt and never return a value). If the expression is not a well-formed program, then an empty string, "", will be returned.

Any BASIC-F `<code>` that is not otherwise well-formed will be interpreted as:

PRINT EVALUATE(<code>)

There is hence no such thing as an ill-formed BASIC-F program.