Repeated quantum suicide would *not* suggest that MWI is correct: a stronger version of the argument.

By Paul Almond

22 February 2011

Website: http://www.paul-almond.com
E-mail: info@paul-almond.com

There is a view that, by performing repeated quantum suicide experiments, an observer can gain evidence that the many-worlds interpretation of quantum mechanics is true. The idea is that, if the observer has survived enough quantum suicide experiments, this would seem implausible if many-worlds were not true and the observer were relying just on "luck", while if many-worlds were true, the observer would know that there would have to be some futures in which he survived – providing an explanation of his current situation. An observer, therefore, should be able to demonstrate to his own satisfaction that many-worlds is true, with any desired degree of confidence, even though onlookers will mainly experience seeing the observer die and the observer can never "come back" to share his knowledge that many-worlds is almost certainly true. In a previous article, *Quantum suicide would not suggest that MWI is correct – even to the person doing it*, available at http://www.paul-almond.com/QuantumSuicide.pdf or http://www.paul-almond.com/QuantumSuicide.doc, the author argued that this view is incorrect and that repeated quantum suicide experiments will indicate nothing about the likelihood of the many-worlds interpretation being correct, even to the observer who is undergoing them. Challenges to this argument are possible based on justification of the choice of reference class of before/after observer moments used and the argument's disagreement with the Bayesian calculation used to show that surviving quantum suicide would suggest that MWI is true. This article provides a stronger version of the argument, which should be less controversial. A reference class of descriptions of possible candidates for your situation at any instant is used, which is not subject to challenges about reference class as the reference class in the previous argument is. The Bayesian calculation used to support the idea that surviving quantum suicide suggests that MWI is true is shown to be a misuse of the Bayesian method.

Table of Contents

1 Introduction 5
2 Possible Objections to the Previous Version of the Argument 7
2.1 Objections About the Choice of Reference Class 7
2.2 Objections Based on the Existing Bayesian Approach 11
3 An Information-Theoretic Consideration: how does surviving quantum suicide affect the ultimate reference class? 16
3.1 The Ultimate Reference Class 16
3.2 The Superiority of the Ultimate Reference Class 16
3.3 The Proportion of You-Centred Possible World Descriptions Corresponding to a Physical Model 17
3.4 Is MWI true? 19
3.5 Surviving quantum suicide does not change the ultimate reference class in any way that matters. 20
3.6 How Use of the Ultimate Reference Class Makes the Argument Stronger 22
4 Viewing Quantum Suicide in Terms of Specificity 24
4.1 Specificity should determine how much of the reference class an outcome gets. 24
4.2 When worlds split in MWI, measure does not magically appear from nowhere: it is conserved. 26
4.3 Reconciling the Reduction in Measure with the Ultimate Reference Class View .. 29
4.4 Surviving quantum suicide would tell you nothing about the likelihood of MWI being true. 30
5 But what about the Bayesian calculation? 35
5.1 The Need to Deal with the Bayesian Calculation 35
5.2 The Bayesian Calculation 36
5.3 The general Bayesian method can be reconciled with the reference class view... 37
5.4 An Example Bayesian Calculation with a Probability Increase 40
5.4.1 An Example of the Bayesian Calculation with Quantum Suicide. 41
5.5 The "Mechanism" By Which a Prior Probability Can Be Increased On Observation of an Outcome 42
5.6 Quantum suicide involves a disconnect between P(Survival) = 1 and any *real* low specificity in the reference class: three views. 45
5.6.1 A Summary of the Three Views 45
5.6.2 View 1: With quantum suicide, making one of the outcomes more likely is being done in a non-standard way that merely removes irrelevant situations from the reference class. 46

5.6.3 View 2: The measures taken in setting up the experiment to result in P(Survival) = 1 are non-standard because they have no special effect on the history of the situations which they are supposed to be making more probable. 48

5.6.4 View 3: With quantum suicide, P(Survival) = 1 involves a non-standard way of assigning probability. 50

6 Some Analogies 52

6.1 Analogy 1: The Clone Factories 52

6.2 Analogy 2: The AI Laboratories 55

6.3 Analogy 3: An Example of How P(Survival) = 1 Can Become Disconnected From Any *Real* Low Specificity in the Reference Class. 57

7 Measure Reduction When Worlds Split in a Level IV Multiverse 64

7.1 How Measure would Decrease in a Level IV Multiverse 64

7.2 Distinguishing Between MWI and non-MWI in a Level IV Multiverse 66

8 Conclusion 70

9 Acknowledgements 73

10 References 74

List of Abbreviations

AI	artificial intelligence
bit	binary digit (0 or 1)
MWI	many-worlds interpretation (of quantum mechanics)

1 Introduction

Quantum suicide is a thought experiment originally proposed by Hans Moravec (Moravec, 1987) and Bruno Marchal (Marchal, 1988), and further developed by Max Tegmark (Tegmark, 1997).

A quantum suicide experiment involves some mechanism that will, depending on some quantum event, either kill the observer or not kill him: we might assume a 0.5 probability of each. The observer can go through a long sequence of such quantum suicide experiments. In the "everyday" view of reality and probability, a long enough sequence of experiments is almost certain to kill the observer: the observer's survival relies on being "lucky" in each experiment, and the more chances that the observer takes, the more likely it is that his luck will eventually run out.

There is, however, another way of looking at quantum suicide. The many-worlds interpretation of quantum mechanics (MWI), otherwise known as the relative state formulation of quantum mechanics and proposed by Hugh Everett, states that when quantum events occur, all possible outcomes are realized. The quantum wave function is viewed as having physical reality, and decoherence causes splitting into separate "worlds" (Everett, 1957; Price, 1995). In this view of reality, the observer in the quantum suicide experiment has multiple futures, and there will always be a future available in which the observer survives.

Max Tegmark has proposed that, if a scientist wants to know whether MWI is true or not, she could use herself as a subject in repeated quantum suicide experiments to find out. The idea is that, after a sufficiently long sequence of experiments, if the scientist is still alive, she should consider it unlikely that this is due to being "lucky" in all the experiments and that, instead, it should be more likely that she is still alive because, in MWI, there is always some future in which she survives. The scientist can supposedly establish that MWI is true with any confidence that she wishes by performing enough quantum suicide experiments: as she survives each latest experiment, the plausibility of a non-MWI explanation decreases and the likelihood of MWI being true increases. In this way, after surviving enough quantum suicide experiments, the scientist can be effectively convinced that MWI is correct. She can never report this back to her colleagues, however: in most of the futures of the scientist's colleagues, the scientist is simply killed early in the process. Quantum suicide is therefore supposed to be a way by which you can establish that MWI is true *just to yourself*.

In an earlier article, *Repeated quantum suicide would not suggest that MWI is correct – even to the person doing it* (available at http://www.paul-almond.com/QuantumSuicide.pdf or http://www.paul-almond.com/QuantumSuicide.doc), the author argued that this idea is incorrect (Almond, 2011). It was argued that, instead, if you have survived any number of repeated quantum suicide experiments, despite what we may intuitively think, this

gives you no evidence that MWI is true that you did not already have. Quantum suicide experiments are of no use in establishing the likelihood that MWI is correct – not even to the person undergoing them.

The argument was based on considerations of probability, observer moments, measure and locating yourself within the MWI multiverse and used a scenario which had some similarity with the Sleeping Beauty scenario (Elga, 2000). The main problem raised by the argument was that of the specificity of observer moments: the observer moments you would enter after surviving a quantum suicide experiment or experiments would be more specific, and you should consider it unlikely that you are experiencing such an observer moment rather than an earlier observer moment – this unlikelihood meaning that you should be just as surprised to find yourself in such an observer moment, rather than an earlier, higher measure, observer moment, as you should be to find that you survived if MWI is not correct.

One way in which the argument in the previous article can be challenged is by questioning its choice of reference class. Why should we group before and after observer moments together statistically? Why not a narrower reference class? Another challenge to the argument can be based on its disagreement with the Bayesian method used to justify the idea that surviving quantum suicide suggests that MWI is true. Is it being claimed that the Bayesian method is fundamentally wrong? Of course, the author cannot reasonably claim this – so we need to show why it is wrong in this context. This article will deal with these objections by presenting a stronger version of the argument, together with an explanation of what is wrong with the Bayesian calculation.

In fact, the issue raised in the previous article is not the main problem with the idea that surviving quantum suicide experiments would tell you anything important: it is, rather, merely a symptom of a deeper, underlying problem: surviving quantum suicide in an MWI reality and a non-MWI reality both result in a reference class of possible situations in which some outcome has happened that is not affected in any important way by any committing of suicide in situations in which that outcome did *not* happen.

Another way of viewing this is in terms of reduction of measure. When quantum events occur in MWI, and before you know the outcome, the situations produced with different outcomes each have lower measure than the original situation from which they split: total measure is conserved and divided up among the situations with the different outcomes. It will be obvious to readers that a similar measure reduction occurs when MWI is *not* true, but as it occurs in MWI, the proportions of situations for MWI and non-MWI versions of some outcome match whatever the proportions of MWI and non-MWI situations were in the reference class before the splitting occurred – and when you find out that some outcome occurred, the situations corresponding to that outcome become the entire reference class, giving a reference class in which the proportion of MWI-situations has not changed. The addition to suicide to the scenario does not change this, as intuition may suggest to some people.

2 Possible Objections to the Previous Version of the Argument

Objections could be made to the previous version of the argument, and the author has been helped in this respect by Max Tegmark, who has asked some challenging questions in an e-mail discussion. Max was not actually disagreeing with the previous argument. He made it clear that he merely wanted to subject the argument to appropriate questioning that would help establish the truth, and help him establish his own position on the subject: he was asking questions as a devil's advocate. Dealing with these challenges has helped in producing this strengthened version of the argument.

2.1 Objections About the Choice of Reference Class

The previous version of the argument made use of a reference class of observer moments before and after quantum suicide experiments. Max asked for justification of the particular reference class of observer moments that was used. There is an issue here of how a reference class is selected: what entitles us to declare that "before" and "after" observer moments are going to be in the same group, with regard to statistics? This question about reference class justification is one that *should* be asked, and the author should have really had an answer to this before Max asked it: it should have been obvious to the author that this would be the weak-point in the argument, where controversy could be generated. In the author's view, the choice of reference class in the previous version of the argument *can* be justified, and it is intuitively obvious that it causes problems for the idea that surviving MWI will tell you anything.

One justification for the previous reference class choice comes from the thought experiment with the "quantum death alarm clock" in the previous article. This thought experiment puts you into a situation in which you go to sleep and are woken either before or after some quantum suicide experiment. In this situation, when you wake up in the thought experiment, and you do not know the date, it should seem just as unlikely that you are in a later observer moment in which you have survived quantum suicide, as opposed to an earlier observer moment, in a reality in which MWI is true, as it is that you have survived in a reality in which MWI is not true. On waking, you do not even know whether you are in a before or after observer moment, and your reference class consists of both kinds of observer moments because of this knowledge. If it is this easy to set up a situation like this, the fact that you would tend to know what the time and date is in a *real* situation, apparently removing this lack of knowledge, should be of little help in getting rid of this reference class. You could imagine that you forget the date for an instant, that your brain takes a very short time to compute it after you survive a quantum suicide experiment, or that you wake up years after surviving a quantum suicide experiment and need a moment to collect your thoughts about what the date is: any of these would put you into a situation in which you should be

subsequently surprised to find yourself in a later, post-quantum-suicide-experiment observer moment if MWI is true.

A further justification for the choice of reference class is that it is quite a conservative reference class. It involves using observer moments at different times for an observer, who remains in the same species of animal and has the same general kind of brain, etc. *The observer moments actually belong to the same person*. This kind of reference class is much less expansive than reference classes routinely used in philosophical arguments such as the Doomsday argument (Carter, 1983; Gott, 1993; Bostrom, 2005).

If an answer is needed to an objection to the choice of reference classed based on the idea that before and after observer moments are somehow profoundly different, we could imagine altering the time at which the quantum suicide events occur in some experiment to turn observer moments into before or after ones arbitrarily. Suppose you take part in a quantum suicide experiment and the actual suicide, if it happens, happens at exactly 10AM. You survive the experiment, and think MWI is more likely to be true, and whatever reference class you use distinguishes between before and after observer moments: you only count the after ones and reject the idea that the before ones should be part of the reference class. At this point, someone walks in and tells you that the actual quantum event and resulting suicide if it was going to occur (which it did not in your world of course) were set to happen at 9:45AM. Observer moments that you thought were "before" ones have just become "after" ones, yet you hardly have any reason to change your view now. The time of the quantum suicide did not really make any important difference to any of the observer moments, except change their measure (and that should be a weak reason for the decision).

A further justification for the choice of reference class is that you should be able to imagine performing a long sequence of many quantum suicide experiments arbitrarily quickly, with an arbitrarily small delay between experiments. This would mean that, for a path through the experiments in which you survive, an observer moment just after the last experiment could be an arbitrarily small time after an observer moment just before the first experiment. Assuming you survive, your brain, and the environment around you would hardly change at all from the first observer moment to the last one, and during all those in-between – and you can make the degree of similarity as strong as you want by doing the entire experimental run more quickly – making a strong case for saying that you can legitimately include all these observer moments in the same reference class.

Do we even need to go as far as doing all the sequence of quantum suicide experiments very quickly, as just described? Quantum suicide as generally viewed involves a single quantum event determining whether you live or die, with the actual suicide, if it is to occur, following as quickly as possible. If we consider an observer moment immediately before this event, and an observer moment immediately after it (in a future in which you survive) this would seem to give us two observer moments which can be arbitrarily close together for a single quantum suicide experiment. The same argument as the one

just used can then be applied: this seems to establish before and after observer moments arbitrarily close together in time on either side of a single quantum suicide event, suggesting that there is nothing significant to distinguish them: the experimental protocol almost seems to have been designed to put the before and after observer moments in the same reference class.

We can also construct a thought experiment that seems to force an advocate of the idea that surviving quantum suicide suggests that MWI is true into dealing with more expansive reference classes than the one proposed here. Suppose we run a quantum suicide experiment that is less clean. Some series of quantum events, will generate a number between 0 and 10^{10}. That number will then determine the extent to which your brain is damaged. If the number is 10^{10} your brain will not be damaged at all. If the number is 10^{10}-1 your brain will be damaged just a little bit. (This hardly matters: your brain could get more random damage in a second of normal life anyway.) If the number is 10^{10}-2 it will be damaged a bit more severely, and so on. The worse case scenario is if the number is 0. In this situation, your brain is damaged to such an extent that it is just a random mess: something that is no more apparently a human brain than it is anything else. The question is: how intelligent do you expect to be after such a process? If you think that P(Survival) = 1 for standard quantum suicide, you should think that some of the really damaged brains that result here should be excluded: they will effectively be dead versions of you, but where do you draw the line, or do you assign probabilities in some variable way? This situation seems to force you into some kind of reference class consideration that is worse than the one presented in the previous argument, so the case here is that the reference class issue in the previous argument is fairly minor in comparison with some more expansive reference classes that need to be dealt with to discuss quantum suicide.

Another thought experiment can be given that should require you to use a reference class even more general than the reference class of before/after observer moments used in the previous version of the argument. The thought experiment is as follows.

A machine is set up to make a human baby using an artificial womb in a laboratory. The laboratory is sealed from the outside world. The human baby will be reared and educated by machines, but its knowledge of the outside world will be limited in various ways.

You were made in this room, and grew up in it. On your 18th birthday, the machines tell you the circumstances of your birth. They were as follows.

One baby was to be made in the room, and allowed to grow up and die of old age, eventually. This is Person 1.

After Person 1 died, a series of quantum events were to be allowed to occur, with a probability of 1 in 1,000 of activating the machinery to make a second human, called Person 2. Person 2, if made, would also be reared and educated in the room.

You are not told whether you are Person 1 or Person 2.

If MWI is not true, you should clearly think that it is unlikely that you are Person 2: there was only a 1 in 1,000 chance that Person 2 would ever be made. Suppose, however, that MWI *is* true. Does this make it plausible that you are Person 2?

Even if MWI is true, you should think it unlikely that you are Person 2. From the argument about the decrease in measure of observer moments in MWI given in the previous article, you should know that Person 2's observer moments will be low-measure and uncommon compared with those of Person 1. Suppose you had to bet on whether you are Person 1 or Person 2? If you think both are equally likely, you have the decrease in measure with which to contend, but if you think it is more likely that you are Person 1, you seem to be accepting a reference class even more general that the one required by the previous version of the argument: Person 1 and Person 2 might be very different people, while the previous argument, in including before and after observer moments in the same reference class, is at least restricting the observer moments to those from the life of a single person – and, as has been described, this can involve observer moments very close together chronologically in that single person's life.

Some justification has been given, then, for the inclusion of before and after observer moments in the same reference class in the previous version of the argument. A problem with this is that there is room to argue with the choice of reference class and these justifications: reference class choices have a tendency to be controversial. The version of the argument given in this article will work in the same general way, but it will use a reference class which should be free of controversy. The reference class used in this article will be the reference class to which the author referred in previous articles as the *ultimate reference class* (Almond, 2011). The ultimate reference class contains every formally expressible description of a possible situation in which you could be, right now, that is consistent with what you know about your situation: it is a set of formally expressed possible world descriptions centred on you as an observer. For example, if you seem to be wearing a red scarf right now, and you remember going to Paris last week, two requirements for any description of a situation in the ultimate reference class are that it is a situation in which you are experiencing wearing a red scarf and it is a situation in which you remember going to Paris last week. The choice of the ultimate reference class does not really need any defence, because it is obvious that every situation in it is a candidate for your situation, and any situation not in it is not a candidate for your situation: it is exactly defined by your knowledge and completely defines the statistics of your situation. For this reason, following the question from Max about the choice of reference class, the author has half-seriously referred to the ultimate reference class in this stronger version of the argument as a "Max-proof" reference class.

2.2 Objections Based on the Existing Bayesian Approach

Max also asked the author to account for the results of the Bayesian calculation used to justify the idea that surviving quantum suicide experiments suggests that MWI is true. This Bayesian calculation is the real, mathematical justification for thinking that surviving quantum suicide experiments tells you anything about physics.

The approach is based on the idea that, before performing a sequence of quantum suicide experiments, you will calculate different probabilities of survival depending on whether you are in an MWI reality or a non-MWI reality. If you are in an MWI-reality, there will always be branches in which you survive – no matter how many quantum suicide experiments are performed – and you can ignore the branches in which you do not survive, as you are not there to make any observations. Therefore, if MWI is true, as far as you are concerned, P(Survival) = 1. If MWI is not true, however, you will not be certain of surviving any single quantum suicide experiment, and as you do more experiments your chance of surviving right to the end becomes progressively smaller, so that P(Survival) can become as low as you want it to be, given a sufficiently long sequence of quantum suicide experiments. If you find yourself alive, after performing a sequence of quantum suicide experiments, you should take into account that your previously computed probability of this outcome (being alive) was higher for the MWI hypothesis and this would increase whatever probability you had previously estimated that MWI is true. If MWI is true, with enough quantum suicide experiments, you should be able to show this with any degree of confidence you want.

The problem presented by this is that this Bayesian calculation does not go away just because someone presents an argument about a reference class of before and after observer moments, as the author did in the previous article. If the previous version of the argument is correct – if surviving quantum suicide tells you nothing about the likelihood that MWI is true – then this is contradicting the Bayesian calculation. Is it really reasonable to question this calculation?

Arguments can be given that suggest there could be a problem with the Bayesian calculation, although such arguments may not indicate exactly what the problem is.

One reason for distrusting the Bayesian calculation is based on its reliance on a particular philosophical assumption about continuity. It is explicitly assumed in the Bayesian calculation that, if MWI is true, you should expect to survive a quantum suicide experiment, or any number of them: that P(Survival) = 1. You are supposed to be able to ignore any branches in which you die, as you are not there to observe anything, and you are supposed to assume that you will always observe a future in which you are alive. This is clearly going to be a controversial idea. In a non-MWI reality things are simpler: the probability that you will survive a particular quantum suicide experiment is the probability of the corresponding outcome occurring, without any consideration of

whether you are there to observe a future in which you do not survive, and as the number of quantum suicide experiments increases, your probability of surviving to the end of the sequence gets closer to 0. Some people, however, do not see it that way, and think that, if MWI is true, this does not justify ignoring the branches in which you die. Instead, you should view each of your possible futures – whether you are alive in it or not – as having some probability of occurring. According to such people P(Survival) = 1 is not justified when MWI is true and, in fact, you should view your chance of survival as being no different than it would be if MWI were not true.

It is not the author's intention, in this article, to get involved in an argument about whether MWI being true implies that P(Survival) = 1. Rather, the problem here is that relying on any philosophical view about continuity is highly suspect: when you find yourself in a given situation, your views about continuity should be irrelevant when deciding how likely some hypothesis is that accounts for that situation.

To see why this is the case, let us consider two scientists, whom we will call Professor Optimist and Professor Pessimist. Both scientists understand how physics should describe reality if MWI is not true. Both also understand MWI, and are open-minded about it and each assigns it the same probability of being true. Where the two scientists disagree is on the issue of continuity. Professor Optimist thinks that, if you perform quantum suicide experiments, you can ignore the branches where you are not there to make observations, so that P(Survival) = 1 if MWI is true. Professor Pessimist, on the other hand, although he thinks MWI might be true, rejects the idea that MWI being true would imply that P(Survival) = 1. He thinks that, if you perform quantum suicide experiments, your chance of survival is just the probability of the corresponding outcome: he thinks that, in a long enough sequence of quantum suicide experiments, P(Survival) will be almost 0, whether MWI is true or not.

Suppose each scientist performs the same, long sequence of quantum suicide experiments and finds himself alive afterwards. How do things look from the point of view of each surviving scientist? If the Bayesian method, in which we look at the previously calculated survival probability, is valid, Professor Optimist should think it very likely that MWI is true, because he could have previously calculated that P(Survival) = 1 for MWI being true, and it would be very low for MWI not being true. Professor Pessimist, on the other hand, should not change his view: he views his survival as unlikely in either case, so his survival tells him nothing.

The problem here is that both Professor Optimist and Professor Pessimist are in the same situation of having survived the sequence of quantum suicide experiments and they have the same views about the nature of reality. They both agree on what reality should be like if MWI is true, and what reality should be like if it is not true. Before the quantum suicide experiments started for each of them, they agreed about the probability that MWI is true, but now that each has survived, their disagreement about continuity of self has not resulted in any disagreement about what the world should be

like if MWI is true. Their previous agreement on the probability of MWI being true was based on what the world would be like to them if MWI is true and if it is not, and nothing has changed in this respect. Professor Optimist's view that P(Survival) = 1, even if it were valid, has not transported him into some radically different universe to that in which Professor Pessimist is, in which MWI takes on some different, more plausible form. Both scientists are in fundamentally the same situation. They should still agree on the probability of MWI being true. This suggests that the Bayesian approach should be questioned.

Another argument that can be made to question the validity of the Bayesian calculation is based on small changes in the experimental procedure. Small changes in the quantum suicide experiments, provided that they do not seriously change the appearance of the world that you end up in if you survive, should not be of any consequence, but the Bayesian argument does make them have consequence. Most people who think that P(Survival) = 1 in an MWI reality have some kind of standard about how the "execution" has to be done to ensure that none of the branches in which you die inconveniently get counted. It is generally required that, if the outcome of the quantum event is that you have entered a branch in which you need to be terminated, you are terminated as quickly as possible to ensure that you never have time to notice that you are on one of the "no survival" branches: that you do not have any time to make an observation in a future in which you do not survive.

Suppose you have set up a sequence of quantum suicide experiments, and you have ensured that your termination, if it were to occur, would happen almost instantaneously after the relevant quantum outcome. The entire sequence of experiments was going to occur with you strapped into a chair which was going to release you automatically when the sequence of experiments is complete, and the entire sequence of experiments was going to occur automatically once started: you were not going to have the option to cancel the sequence before it completed. You have just survived the sequence of quantum suicide experiments, and you are thinking about whether or not MWI is true. You remember thinking, before the experiments started, that the termination process was going to be very quick and meet whatever standards you have, and therefore you feel justified in thinking that P(Survival) = 1. You go through the familiar Bayesian reasoning to conclude that MWI is very likely to be true.

At this point, someone walks into the room, and says, "Sorry! We tampered with your experiment and played a practical joke on you. You *thought* that the termination process was going to be so quick that you would never experience the non-survival branches, but we changed your experimental apparatus so that if the quantum event outcome indicated you were going to be killed, a sign lit up in front of you saying 'You are going to die in 5 minutes' and you would be given five minutes to think about that, strapped in the chair and unable to cancel the process, before you were killed. That means that, if MWI is true, when you entered branches in which you were going to die, you knew this had happened. If you had known this before starting the experiments,

you would never have thought that P(Survival) = 1 for MWI being true: even if you had known that MWI *is* true, you would have been worried about the possibility of entering one of those non-survival branches, aware that if it happened you would experience it and it would be just as real, together with your impending death, as one of the survival branches. You never knew about this tampering with your experiment until now because, of course, your history is that of someone who has survived, and so you have never seen what happens when the quantum event outcome indicates non-survival." Suppose this person gives you adequate evidence to persuade you that he is telling the truth. He may show you the apparatus that was tampered with, witness statements of people saying they saw the experiment being tampered with, etc: whatever he provides is enough to convince you.

You now know that, because of the tampering, you were wrong to think that P(Survival) = 1 if MWI is true, so if your view that MWI is probably true was based on that, you should now adjust this view: the Bayesian argument is of no use now for supporting MWI. However, it should make little sense to adjust your estimate of the probability that MWI is true on the grounds that, if it is true, some other versions of you were mistreated for five minutes. *None of this has any important effect on the situation that you are in right now*. Your current situation is almost the same as it was before: it just happens to be the case that if you are in an MWI multiverse, some versions of you had a disappointing experience. This does not significantly affect the situation that you are in if MWI is true. The observer moment you are in right now is fundamentally the same observer moment, and you got into it in fundamentally the same way. Yes, this five minute horrible experience for the other versions of you might make a difference to continuity and whether you should have expected to survive, but it is unrealistic to think that it can dramatically change the characteristics of *your current situation*. Whatever method you use to decide on the probability that MWI is correct should be based on the characteristics of your situation in an MWI multiverse and the characteristics of your situation in a non-MWI reality – and nothing has changed here. If MWI is true, you have basically the same history that you would have had without the tampering with the experiment, because the main effects of the tampering – the frightening five minute wait for the version of you doomed to die – is something that has never been encountered in your history. How then, can you reasonably think that your view of the probability of one description of your situation – MWI – being correct versus the probability another – non-MWI – being correct can have changed due to events that have not even occurred in your past or done significant to affect the world around you? What you should actually be considering is the *specificity* of your situation in an MWI-reality and a non-MWI reality – which cannot have changed significantly at all. This completely contradicts the idea that seemed to follow from use of the Bayesian hypothesis that the tampering with the experiment does require you to change your probability for the MWI hypothesis. All this should bring the Bayesian approach into question.

The two arguments just given should have brought into question the idea of using a Bayesian calculation to justify the idea that surviving quantum suicide experiments tells you that MWI is true, but neither argument actually dealt with the Bayesian argument directly, and some answer is needed. It is not enough to say that the Bayesian approach is invalid: we must show *why* it is invalid. The author could attempt to show that the kind of Bayesian approach used here is invalid in general. This would be an extreme position to take, given the great success of the Bayesian approach, and would probably mean that the author is doing "crank philosophy". The author will *not* be taking this position and is making no attempt to get rid of the general Bayesian method. Instead, in the stronger version of the argument in this article, it will be shown that use of the Bayesian approach in this kind of situation – using it to infer that MWI is a more likely hypothesis, given survival of quantum suicide experiments and an assumption that P(Survival) = 1 in MWI – is invalid – that this kind of use of the Bayesian approach has an unusual feature that makes it a *misuse* of the Bayesian approach. The idea that there are two legitimate ways of dealing with quantum suicide survival – using reference classes of observer moments and applying the Bayesian approach, based on the assumption that P(Survival) = 1 for MWI being true – with each approach having valid arguments that can be used to justify it, and with there being some reasonable controversy over which approach to use, will be shown to be wrong. It is not just a matter of which view we prefer – Bayesian statistics or reference classes of observer moments: thinking that surviving quantum suicide experiments tells you anything about the likelihood of MWI being true is *wrong*.

With the stronger version of the argument that will be presented involving a much more tightly defined reference class that is safe from reasonable controversy and with the Bayesian justification being explicitly shown to be invalid, there will be no reason left to think that surviving quantum suicide will tell you anything about the likelihood of MWI being true.

3 An Information-Theoretic Consideration: how does surviving quantum suicide affect the ultimate reference class?

3.1 The Ultimate Reference Class

In previous articles, the author gave the term *ultimate reference class* to the set of all possible descriptions of your situation – all possible situations in which you could be right now (Almond, 2011). Each possible situation in the ultimate reference class is a you-centred possible world description. You-centred means that to qualify for entry in the ultimate reference class, a possible world description has to describe a candidate for your situation right now. The entire description is centred on you. It has to be consistent with everything you know. e.g. if you remember wearing a red necktie last Tuesday, every member of the ultimate reference class – every you-centred possible world description – must be for a version of you that remembers wearing a red necktie last Tuesday. Basically, every member of the ultimate reference class is a candidate for your situation right now, and it contains every possible candidate for your situation.

The ultimate reference class will contain an infinity of you-centred possible worlds, but we will deal with this by limiting the description length to n binary digits (bits). However, you can let n tend to infinity to consider the statistics of your situation. The ultimate reference class defines the statistics of your possible situation and the probability that you are in some kind of situation – a situation with some property – depends on the proportion of situations in the ultimate reference class with that property. For example, if you want to know the probability that it will snow outside your window tomorrow then, *in principle*, you can compute the ultimate reference class for some very large value of n, and determine the fraction of you-centred possible world descriptions in it for which this is the case: that is the probability that you are in a situation in which the snow will arrive.

3.2 The Superiority of the Ultimate Reference Class

The ultimate reference class should be regarded as superior to all other arguments about the world around you and what it is doing or likely to do. That does not mean that other arguments are wrong: it merely means that any argument that tells you about expectations can only validly work by telling you about the ultimate reference class's likely structure. If an argument disagrees with the ultimate reference class when the ultimate reference class is computed for some value of n, then either n is too low – and the ultimate reference class will start agreeing as n is raised – or the argument is simply wrong. The ultimate reference class must always win, because it completely describes

the statistics of your situation in a way that is free of any prejudice. Anything that could not possibly be your situation is not in there. Anything that could be your situation is. We do not pick and choose between members of the ultimate reference class, but instead we treat them all equally. However, this does not mean that all claims about the world are the same: some claims will correspond to a greater proportion of you-centred possible world descriptions in the ultimate reference class than others.

3.3 The Proportion of You-Centred Possible World Descriptions Corresponding to a Physical Model

A you-centred possible world description, for some particular scientific model, when there is a maximum description length of n bits, will contain the following information:

- some information, **G**, that describes the general physics of that model. As an example, a you-centred possible world description in a Newtonian reality would need information to describe Newton's laws.
- some information, **K**, that describes what you know about your situation that distinguishes it from other observer-centred possible world descriptions that have the same general physics, G, but are not consistent with your situation. As an example, a you-centred possible world description in a Newtonian reality would need information to describe the positions and velocities of particles *that you know about*. Any observer-centred possible world description with this particular G needs this particular K to be a candidate for your situation – to be a you-centred possible world description.
- some information, **S**, which is just extra information about the situation that does not necessarily relate to your actual situation, but could do. S is different in different you-centred possible worlds. As G and K must be identical for all you-centred possible worlds with the same general physics model, they can only be different by having different S. As an example, a you-centred possible world description in a Newtonian reality would need information to describe the positions of velocities of particles *that you do not know about*, or it could provide extra information about the positions and velocities of particles that you do know about, but with more accuracy than any measurements you have made.

Each MWI or non-MWI you-centred possible world description, for some maximum description length of n bits, therefore contains some information, G, some information, K, and some information, S, with the total amount of information for G, K and S not exceeding n bits. Not every observer-centred possible world description that can be made like this would actually be a *you*-centred possible world description. To be a *you*-centred possible world description, the observer situation that is described must be consistent with what you know about the specifics of your situation, and K must match this. For example, if you have the perception of living in a house with a green door, the bits in K will need to reflect this in any you-centred possible world description for that G, so the same K will be needed for all you-centred possible worlds for the same G. Any K

which is for an observer who is not having the perception of living in a house with a green door is part of someone else's observer-centred possible world description, but it cannot be part of a you-centred possible world description.

The ultimate reference class, with some maximum description length of n bits, has some proportion of MWI-type you-centred possible world descriptions and some proportion of non-MWI-type you-centred possible world descriptions. Each you-centred possible world description must be unique. All the observer-centred possible worlds conforming to the same general physics will have the same G, and so two observer-centred possible worlds with the same general physics, G, can only be different by having different known and/or unknown information for the observer's specific situation, K and S. K must be the same for all *you*-centred possible worlds with the same G, however, so two you-centred possible worlds with the same general physics, G, can only be different by having different S. The total amount of information for G, K and S, however, must not exceed n bits, so the more bits of the description that are used up by G and K, the fewer are the bits that are left over for S, and the less is the number of different you-centred possible world descriptions that can be resolved apart with this lower number of bits.

As extreme examples:

- If n were 100 bits, and G for some physical model needed 60 bits to express it, and 25 bits were needed to express K – what you perceive to be the case about *your* situation with that physical model – then this leaves 100-60-25=15 bits for expressing different versions of S and resolving you-centred possible world descriptions apart for that general physical model. The number of you-centred possible worlds with this maximum description length of 100 bits and this general physics model is therefore 2^{15} = 32,768.
- On the other hand, if n were 100 bits again, and G for some physical model needed only *30* bits to express it, and 25 bits were still needed for K, this would leave 100-30-25=*45* bits for expressing S and resolving you-centred possible world descriptions with that general physics model apart. The number of you-centred possible worlds with this maximum description length of 100 bits and this general physics model is therefore $2^{45} = 3.5 \times 10^{13}$. This physical model would be much more likely to correspond to your actual situation than the first one, because you-centred possible world descriptions corresponding to this model are taking up much more of the ultimate reference class.

All else being equal, the amount of information needed to express G for some physical model, in comparison with the amount of information needed to express G for alternative physical models, is what determines the probability that a physical model is correct. If you want to know the probability that you are in an MWI-type reality – that MWI is true – you should estimate the amount of information needed to express G for MWI versus the amount of information needed to express G for a typical non-MWI physical model, and this will tell you the relative numbers of you-centred possible

worlds that can be made for each and therefore the relative proportions of each in the ultimate reference class.

There is a minor complication here: some of the bits used to express the observer-centred possible world description are needed for K, the known specifics of your situation, to distinguish your situation from those of other observers who could be imagined in situations with that same general physical model. In the above example, it was assumed to be the same in each case – and this assumption will usually be reasonable. An exception is when you find a general physical model, G, that actually explains some of the specificity in your situation as being an unavoidable part of the general physics: this would tend to reduce the number of bits needed for K and increase the probability that G is correct – unless G is itself merely a very specific physical model out of many similar possibilities, with the bits that have been taken out of K effectively having been put into it.

3.4 Is MWI true?

There is a common view that we can never know about the chance that a theory such as MWI is true – that, because both MWI being and MWI not being true may seem to fit observation equally well, the issue is somehow "beyond the horizon" of things that we can know about and equivalent to "How many angels on the head of a pin?" type questions, but reasoning like that just given in 3.3 suggests that this view is wrong. All theories are *not* equal – and two theories are not equal even when they both fit the observations. If one theory needs less information to express G than another theory then, all else being equal, it will be more likely that this theory is correct. Furthermore, if you can actually get the computing power to compute the reference class you can, at least in principle, put numbers on the probabilities.

G for both MWI and non-MWI realities will need the physics of how the wave function works. However, G for a non-MWI reality will need some extra information: some mechanism for wave function collapse will need describing – and this is not needed in MWI as wave function collapse is not part of the model. This should mean that less information is needed for G in an MWI reality than in a non-MWI one, meaning that more unused bits are left over for K and S in an MWI reality. Unless you can find one (and it will be harder than many people might think), there is no good reason to think that the amount of information needed for K is any different for MWI or non-MWI physical models. You should think that more unique you-centred possible worlds can be described for MWI and they should occupy more of your ultimate reference class. *This should suggest that MWI is probably true*. A common objection to MWI is that the massive number of worlds is absurdly unparsimonious, but the above reasoning should show that this is a misunderstanding of how we should view parsimony. *Parsimony should mean less information for G and K*. For a given observer, all the other worlds that are "out there", relative to his own situation may seem to need a huge amount of information to describe them, but this is irrelevant. When you are making an observer-

centred possible world description, with some maximum description length of n bits, you do not need to describe all these other worlds: we must merely say enough about this version of your situation to distinguish it from all the other situations in the ultimate reference class with that value of n. One way of viewing this is by saying that you should have a world view which has as few "entities" as possible, so that you are not "multiplying entities beyond necessity" *and that each bit of information in the description is an entity*. The author suggests, on the basis of all this, that MWI is probably true. Eliezer Yudkowsky has previously argued for MWI being true on the basis of information content in the theory (Yudkowsky, 2008). The likelihood of MWI being true, however, is not the main subject of this article: the issue being discussed is whether surviving quantum suicide experiments tells you anything about the probability that MWI is true, so we will now consider what surviving quantum suicide experiments does to your ultimate reference class.

3.5 Surviving quantum suicide does not change the ultimate reference class in any way that matters.

How does performing and surviving quantum suicide experiments affect all this? *It should be apparent that surviving quantum suicide is not going to have any effect on the relative proportions of MWI and non-MWI you-centred possible worlds in the ultimate reference class*. Suppose that before performing any quantum suicide experiments there was some value of the amount of information in G for MWI being true and some value for the amount of information in G for MWI not being true. There would also be values for the amount of information in K, the known information about your specific situation given the relevant physics model (MWI or non-MWI) – and we should expect the amount of information needed for K to be the same in each case, but we need not assume this when considering quantum suicide.

These values would indicate how much information is needed to describe the general physics for an observer-centred possible world (G), together with those features of the situation that are specific to your situation and known to you (K), and for some maximum description length of n bits, how many bits would be "left over" for S to describe a unique situation for you. The amounts of information in G and K (though we can expect K to be the same) for MWI being true and false should therefore determine the probability that you assign to MWI being true. In practice, you may not have properly calculated numbers of bits to work with, but you may have some rough idea of the proportion of the reference class that you expect a type of situation to occupy. (If you estimate a probability that MWI is true, before any experiments are performed, this is what it really means: you can imagine it as making an estimate about the behaviour of a computer programmed to work out the ultimate reference class for some large value of n, with n tending to infinity.)

Suppose that you now perform some sequence of quantum suicide experiments and survive. How does this change things? In terms of the proportions of the reference class

taken up by MWI and non-MWI observer-centred possible world, nothing has changed at all. G, the information needed to describe the general physics, remains the same for MWI and non-MWI situations, and for each the number of you-centred possible worlds is still dependent on the number of bits not used by G and K and available for S to specify different observer-specific situations for you.

There is a minor complication here. There may not be a single value of G for MWI-situations: we might imagine different ways in which the physics of an observer-centred possible world with MWI could be specified. Likewise, we may imagine different ways in which physics could work without MWI. This is not a problem. We can use the same reasoning for any particular G. For some particular G, the information describing the general physics in an observer-centred possible world, whether it is an MWI or a non-MWI possible world, there will be a set of observer-centred possible worlds with that G – with the same general physics – and the number of such you-centred possible worlds, for some maximum description length, n, will depend on how many bits are available for S to describe them. This will depend on how many bits are "left over" after the bits have been used up by G and the relevant K, implied by that G, for your situation. This will be the case both before and after surviving quantum suicide, so the number of observer-centred possible worlds corresponding to the same general physics, described by some G, should remain the same. We can apply this reasoning for every different G that was in the ultimate reference class before the quantum suicide experiments were performed to see that the number of observer-centred possible worlds in the ultimate reference class after the quantum suicide experiments have been performed is unaltered.

Surviving quantum suicide experiments would have no important effect on the proportions of you-centred possible worlds in the ultimate reference class for each general physical model, and it would therefore have no effect on the probability that a particular theory is true.

Some readers may object to this by saying that this is ignoring what we find out from the experiments: that we have survived. To show why this does not work, a consideration of it will be given.

Suppose you have not yet performed any quantum suicide experiments and you compute the ultimate reference class for some maximum description length, n. The proportion of MWI and non-MWI you-centred possible worlds in each case will depend on the number of bits needed in G and K, with the "left over" bits being available for specifying different observer-centred possible worlds.

Suppose now that you perform ten quantum suicide experiments, each with a possible outcome of 0 (death) or 1 (survival). Each time you perform an experiment, and survive, you know that you survived and that the result of that experiment was 1. When you know that you have performed all ten experiments and survived, you know that the result was 1 in each case, and therefore that the results for all ten experiments were

1111111111. Suppose you think that this information is part of the description of your situation, and needs to feature in any observer-centred possible world description. The information about the experiment results is specific to your situation as an observer, and so would be part of K, the information specific to an observer's situation. It would take up ten bits, and these will all be the same for all you-centred possible worlds in the ultimate reference class: they are not available for S and cannot be used for resolving observer-centred possible worlds apart. The result of this is that the number of bits available for making different versions of S has effectively been reduced by ten. Consider some G, corresponding to some general physics model, with a maximum description length of n bits. Before starting the quantum suicide experiments, there will be a number of you-centred possible world descriptions in the ultimate reference class corresponding to G. For each of these, some of the available n bits are used up by G and K, and the remaining bits are available for S – and the number of unique observer-centred possible worlds that can be made depends on how many of the n available bits are left for S after the ones needed for G and K have been used up. Now consider the same G after you have survived the ten quantum suicide experiments. If you now require that K for each observer-centred possible world that uses G also contains the information that the experimental results for all ten experiments were 1111111111, then the number of bits available for S for making unique you-centred possible world descriptions corresponding to G has been decreased, so there will be fewer such you-centred possible worlds in that reference class. However, this can be said for any G – and none of this has assumed that G is MWI-physics or non-MWI-physics: it applies for any G regardless of whether G is an MWI-model or a non-MWI one. The result of having to put the ten extra bits to describe the experimental results into K for each possible world description affects the numbers of observer-centred possible worlds in the ultimate reference class for both MWI and non-MWI models equally. Overall, there is no change in the *proportion* of you-centred possible worlds in the ultimate reference class in which MWI is true.

In fact, it would be absurd if this were not the case. Your situation continually becomes more specific: you acquire information about reality all the time, and if any of it altered the proportions of observer-centred possible worlds for different physical models in the ultimate reference class you would not even need to consider performing quantum suicide experiments. Just living, and watching time pass and things happen would presumably give you enough information to change your ultimate reference class to favour one view or another, and this is clearly not the case.

3.6 How Use of the Ultimate Reference Class Makes the Argument Stronger

The version of the argument being given here may seem to be very different to the one given in the previous article, but the previous argument is really about the same issue. Quantum suicide cannot tell you whether MWI is likely to be true because, when you

find you have survived, you should expect situations in which MWI is true and that outcome has occurred to have the same representation, relative to that of situations in which MWI is not true and the same outcome has occurred, that MWI and non-MWI situations previously had relative to each other in your reference class. This can also be viewed in terms of specificity: as quantum events occur, situations with particular outcomes become increasingly specific for both MWI and non-MWI scenarios, with neither gaining an advantage. It can also be viewed in terms of measure reduction: the measure of observer moments decreases when worlds split in MWI (as long as you have not yet found out what the outcome was), and this applies to both MWI and non-MWI scenarios. When you know the outcome, your new reference class is made from both MWI and non-MWI versions of situations with that outcome, which have already had their measures reduced equally.

The previous argument showed how the measure reduction deprives you of information by using a reference class of observer moments before and after the quantum suicide experiments to suggest that you should consider it just as unlikely to find yourself in one of these later "survival" observer moments as an observer moment before the quantum suicide experiments. Surviving quantum suicide, if MWI is true, should still mean that you are in a situation that is unusual, and this should need as much explaining as the luck you will seem to have experienced if you survive quantum suicide if MWI is not true.

In discussions with Max Tegmark, he pointed out that the author was using a reference class which may need some justification: the reference class of observer moments before and after the quantum suicide experiment. Max asked for reasons for choosing the reference class: why not one which was narrower, or wider? Justification for the choice of that reference class has actually been given in 2.1, but the stronger version of the argument that has been given now does not rely on such justification, because the choice of reference class is clearly justified. The ultimate reference class is the set of all possible descriptions of situations which are consistent with what you know about your situation right now, subject to some maximum description length of n bits. This reference class fully defines the statistics of your possible situation.

4 Viewing Quantum Suicide in Terms of Specificity

4.1 Specificity should determine how much of the reference class an outcome gets.

An argument has just been given in Section 3 showing that surviving quantum suicide tells you nothing about the likelihood of MWI being true if you consider your situation as corresponding to an observer-centred possible world in the ultimate reference class – the set of all observer-centred possible worlds that are consistent with what you know about your situation, for some maximum description length, n, as n tends to infinity. The idea of the ultimate reference class being viewed in information-theoretic terms like this may be controversial, so an alternative view of the same kind of argument will now be presented that avoids information-theoretic type language of "bits", and instead considers things with regard to specificity. The argument is essentially the same: specificity is really the main issue here – whether that specificity is considered in terms of information bits or in some more simplistic idea of the number of ways in which a situation can arise, as will be the case here. To gain an understanding of things, we should go right back to basic probability. Now, the author is well aware that most people reading this will be familiar with the concepts of basic probability, and is not attempting to educate people about it. Rather, we need to revisit it to see what it tells us about surviving quantum suicide.

When we perform probability calculations, we are used to the idea that the greater the number of ways in which an outcome can occur, the more likely that outcome is. An example familiar to most people reading this will be that of rolling a pair of dice. Suppose you have two dice, each with six faces numbered 1-6, and you are going to roll them and add the two numbers together. There are four different ways in which you can get a total of 5: 1+4, 2+3, 3+2 and 4+1. However, for a total of 3, there are just two ways: 1+2, 2+1. There is only way of getting a total of 2: 1+1. The fewer the ways in which a particular total can be attained, the more specific that total is, while the greater the number of ways it can be attained, the less specific it is.

When we do not know what outcome to expect, or when an outcome has occurred but we do not know what it was, we view the less specific kinds of outcomes as more likely. We are effectively making a reference class of possible situations, with each way a kind of outcome could occur being a possible situation in that reference class. The number of ways in which a kinds of outcome can occur is the number of situations in the reference class corresponding to that kind of outcome, so the less specific kinds of outcome occupy more of the reference class. If you do not know what your situation is going to be, or what it is now, you assume that it is randomly selected from the reference class

of possible situations, and so the specificity of a situation indicates how likely it is that that situation is yours, or is going to be yours.

This may seem different to the idea of thinking in terms of the ultimate reference class of formally expressed descriptions of observer-centred possible worlds, previously discussed, but it is essentially the same idea. With the dice example, instead of thinking in terms of unique descriptions, we are considering one aspect of the description: the numbers that appear on the dice, and we are treating each pair of numbers as a different situation, so that 2 on one dice and 3 on the other, for example, counts as one situation. In reality, there will be many ways that each pair of numbers can be obtained: the dice could move in slightly different ways, or land in slightly different places and still give the same numbers, but when we are only interested in the dice, the simplification of assuming that the dice results represent the entire description of the world is entirely reasonable. There may be many different ways in which 1+3 can be obtained with the dice, and many different ways in which 2+3 can be obtained, but we can assume that it is the same in each case – that each pair of numbers corresponds to the same, very large number of situations in which that pair appears. Each of these situations would actually be a formally expressed observer-centred possible world description in the ultimate reference class. When we are doing the probability calculations for the dice, the assumption that each pair of numbers corresponds to the same number of possible situations allows us to treat each pair as being a single situation for the purpose of comparing them with each other.

For example:

For a total of 5 with two dice, there are 4 possibilities: 1+4, 2+3, 3+2 and 4+1. If we say that each pair of numbers corresponds to w possible situations, then this corresponds to 4w situations in the reference class of possible situations.

For a total of a total of 3, there are just 2 possibilities: 1+2 and 2+1. If we say that each pair of numbers corresponds to w possible situations, then this corresponds to 2w situations in the reference class of possible situations.

The number of situations where the total is 5, relative to the number of situations where it is 3, will be $^{4w}/_{2w} = 2$ – and w cancels. For simplicity, we can treat w as if it is 1, and as if each pair of dice numbers corresponds to a single situation.

Similarly, the total number of possibilities for the two dice is 6x6=36, so if each pair of numbers corresponds to w situations, the total number of possible outcomes is 36w. The probability that a total of 5 is obtained is $^{4w}/_{36w} = {}^{1}/_{9}$, and again, we can treat the situation as if w were 1.

In the discussion about the ultimate reference class, we considered things in terms of descriptions of situations, while it may seem that, here, we are looking at the history of a situation rather than its description, but there is no real difference. The history of a

situation is part of the description of how things are, and if two situations have different histories they will be different situations. Suppose we roll a pair of dice and do not look at the result. Instead, we hide the dice under a cloth and we want to know the probabilities of various totals being given by the dice under the cloth. The specificity of each total tells us how many ways that total can be obtained, assuming that each outcome corresponds to one situation, but as has been shown, this is just a simplification: it is telling us about the relative numbers of situations in the ultimate reference class. Each different way of obtaining a total corresponds to a situation with a different description of how the past appears to an observer in that situation, so the total on the dice is telling us about the number of ways in which one feature of the situation – the past behaviour of the dice with regard to the pair of numbers that have resulted – can be different. When we are thinking in terms of probability of different totals on dice we are dealing with probability in a simple, restrictive way. In reality, situations in the ultimate reference class of possible situations could be different in other ways. For example, we might imagine the same outcome, in terms of numbers on the dice, occurring in possible situations in which MWI is true or not true, in which laws of physics are different in other ways, or in which the basic laws are the same but the physics constants are different. In fact, we might reasonably argue that there is an infinity of possible situations, which is where the approach used with the ultimate reference class of having a finite description length of n bits, which tends to infinity, comes in. The main idea, though, is that the probability that you are in a particular kind of situation should depend on the extent to which it is represented in your reference class of possible situations. This in turn depends on the specificity of that kind of situation: the more specific a situation is, the less common it will be in the reference class. When considering things like dice rolls, and the probabilities of obtaining various results, all this can often be simplified and we can assume that, when the situation is described in some way, each outcome is equally represented in the reference class and a kind of situation with low-specificity is one associated with many different outcomes.

4.2 When worlds split in MWI, measure does not magically appear from nowhere: it is conserved.

The next important idea is that *total measure is conserved when worlds split in MWI*: measure does not magically appear when worlds split, but instead it is shared out between the worlds that have split off. In the author's previous article on this subject, this was shown to be the case using an argument based on "Doomsday argument" type reasoning. It was argued that, if a world splits into two worlds, each with an observer moment immediately after the split with the same measure, and if each new observer moment had the same measure as the old observer moment in the world before the split, this would mean that the proliferation of worlds would be accompanied by a proliferation in measure of observer moments: if we treat all observer moments equally, for the purpose of statistics, almost all the observer moments in your life would be right at the end of it, so you should expect to be experiencing an observer moment

now which is a tiny fraction of a second before your death – or an observer moment even later in the universe in the last fraction of a second of someone else's life. The fact that you are not experiencing any such observer moment right now means that the total measure of observer moments at any time, assuming conventional survival, should be about the same, implying that measure is reduced when worlds split.

The author, although finding this persuasive enough, would prefer not to have to rely on "Doomsday argument" type reasoning in this article. The idea now is to avoid possible controversy about reference class choice. This has already been done in this article by using a reference class consisting only of possible candidates for your situation right now. The "Doomsday argument" type justification for the idea that measure is reduced when worlds split in MWI could be subject to the same kind of controversy, so a better argument is needed. A better justification for this is as follows.

Suppose, prior to performing a sequence of experiments, you make some estimate of the probability that MWI is true. You should really think of this in terms of reference class: it should mean that you can imagine a reference class of possible situations – or you-centred possible world descriptions – in which you could be, right now, and your probability of MWI being true corresponds to the proportion of situations in your reference class in which MWI is true. How you make the estimate is up to you: you can apply whatever criteria you normally apply to assess the probability of a theory being true.

You start a sequence of experiments. In each experiment, a quantum event occurs with the result 0 or 1, which is recorded. This is not a quantum suicide experiment: there is nothing set up here to kill you and you will always survive regardless of which result occurs. You do not get to see the result of each experiment. Instead, it is printed out on a printer that is in a locked cupboard, so after each experiment, you know that either 0 or 1 was printed, but without opening the cupboard you do not know which.

You perform 1,000 experiments like this, with the cupboard remaining locked all the time. At the end of the sequence of experiments you know that the printout in the cupboard displays a list of 1,000 1s and 0s – but you do not know what the individual results are.

Let us consider how your probability distribution of possible situations should have changed during the sequence of experiments. After the first experiment, you know that either 0 or 1 had been printed out, but you did not know which. You know that your reference class of possible situations should contain some situations in which MWI is true and 0 was printed out, some situations in which MWI is true and 1 was printed out, some situations in which MWI is false and 0 was printed out and some situations in which MWI is false and 1 was printed out. Suppose that the proportions of situations in which MWI is true and 0 was printed out and MWI is false and 1 was printed out, added together give an amount that is greater than the proportion of the reference class

corresponding to MWI being true, before the experiment was performed and before the splitting of the world associated with the experiment occurred. This would mean that the total proportion of the reference class corresponding to MWI had just increased: MWI would have become more likely to be true. Further, if this happens for one experiment it should happen for each of the 1,000 experiments in the sequence. Every time an experiment in the sequence is performed, the proportion of situations in the reference class corresponding to MWI should increase and the probability that MWI is true should increase. Performing a long enough sequence of experiments should satisfy you, with any degree of confidence that you want, that MWI is true. This, however, would be a strange result. Nothing of any significance is happening here. A sequence of quantum events is occurring with a 0 or 1 result, but quantum events happen all the time. If the probability of MWI being true should increase just because 1,000 quantum events have caused the printing of 1,000 0s and/or 1s in a cupboard, the huge number of quantum events occurring naturally all the time should increase the probability that MWI is true without us having to do any experiments at all. If we just wait while quantum events occur around us, our reference class should become increasingly dominated by situations in which MWI is true and MWI should be effectively proven to us just by waiting around and doing nothing: in fact, it should have been proven to us long since because, presumably, all the quantum events that occurred before we were born should have filled up our reference class with MWI situations.

This result is clearly absurd, and we should not take it seriously. The proportion of the reference class occupied by MWI-situations should clearly stay the same in experiments like this – leaving quantum suicide aside for now – and this means that any splitting of worlds in MWI must correspond to measure being conserved and divided out among worlds.

There is nothing special about MWI in this respect. Of course, the proportion of the reference class occupied by the kind of situation in which MWI is not true should also be conserved during a sequence of experiments like this, and this means that when an event can have a number of outcomes, the proportion of the reference class occupied by non-MWI situations must be divided out among the different outcomes, so that each occurrence of a specific outcome in one of the experiments in the sequence corresponds to a reduction in measure. Of course, when you find out the outcome, the situations corresponding to that outcome, which have just been given a share of the measure, take over the entire reference class: this increase in measure is not anything profound, but can rather be considered as a kind of "renormalization".

The idea that splitting of worlds causes measure reduction, before you know the outcome, can also be justified just by thinking in terms of specificity. Suppose, before the experiment, we consider some situation in the reference class corresponding to MWI being true in which some, or none, of the sequence of 0s and/or 1s has already occurred. When we know that the experiment has occurred, but before we know the result, we know that our reference class now has a version of that situation in which the

outcome was 0 and a version in which the outcome was 1. (Actually, there will be many versions of each, but we do not need that complication.) The same, however, can be said for any situation in the reference class before the experiment: when the experiment occurs, every situation in the reference class gets replaced by 0 and 1 versions. Because this happens in the same way with every situation, the proportion of the reference class occupied by the replacements of a situation, with 0 and 1 outcomes, after the experiment, is the same as the proportion of the reference class occupied by that situation before the experiment. The proportion of the reference class occupied by some situation must therefore be divided up among the 0 and 1 versions of it that exist in the reference class after the experiment. Each time a world splits, the situation becomes more specific, so that, before you know the outcome, each outcome gets a smaller proportion of the reference class. Again, this applies to the non-MWI case as well: each outcome is more specific than the previous situation from which it is derived, causing the measure to be divided up among the different outcomes.

4.3 Reconciling the Reduction in Measure with the Ultimate Reference Class View

All of the discussion just given in 4.2 is consistent with the previous discussion of the ultimate reference class in Section 3. If we consider the ultimate reference class of you-centred possible worlds with some maximum description length of n bits, before you perform one of the experiments in the sequence, each of the you-centred possible world descriptions in which MWI is true contains some information, G, describing a general physics model for a reality in which MWI is true, and some information, K, describing what is known to you about your situation given that general physics model. There is other information, S, in the description, describing your specific situation. For any G, it is the information in S that will resolve two you-centred possible world descriptions apart and allow them to exist as unique descriptions in the reference class. For a situation in which MWI is true, the experiment has now occurred and 0 was printed in the cupboard, the fact that 0 has occurred is extra information about your situation. One bit of information in S is needed to indicate that 0 was printed in the cupboard. Using up a bit in S means that one less bit is available to resolve different worlds apart: it halves the number of unique descriptions that can be made. Similarly, for a situation in which MWI is true, the experiment has now occurred and 1 was printed in the cupboard, one bit of information is needed in S to describe this. Again, this means that one bit less is available in S to resolve descriptions apart, halving the number of unique descriptions that can be made. All this should make sense if we consider that the number of bits available for S is the same before and after the experiment, so the number of you-centred possible worlds with some G must remain the same and be divided up among the different outcomes. Again, there is nothing special about MWI in this respect. The same applies when G is a non-MWI general physics model: the total measure will remain the same and will be divided up among the different outcomes, and this can be justified with the same reasoning. It should be

noted that, although one world in which an outcome occurs with two possible outcomes, 0 or 1, gives rise to two worlds, if the description lengths exceed n, the number of you-centred possible worlds for MWI or non-MWI does not increase for that value of n: the maximum number of bits imposes a limit on the number of you-centred possible world descriptions that can be made.

4.4 Surviving quantum suicide would tell you nothing about the likelihood of MWI being true.

It has been shown that when we consider a sequence of experiments, each involving a quantum event with an outcome of 0 or 1, with the outcomes of the experiments being hidden from you, while you are unaware of the outcomes, the proportion of MWI and non-MWI situations in the reference class does not change. Instead, as each experiment occurs, each situation in the reference class is replaced by 0 and 1 outcome versions of that situation, with the measure being divided up between them: the proportion of MWI or non-MWI situations remains the same.

Suppose that you have gone through all 1,000 experiments, and you have still not opened the cupboard. You know that the cupboard contains a printout with a list of 1,000 0s and/or 1s on it. There are $2^{1,000}$ possibilities for the sequence of 0s and/or 1s on the printout in the cupboard. The probability of any particular sequence being the one on the printout in the cupboard is extremely unlikely. For any particular sequence there will be many corresponding situations in the reference class. Some of these will be situations in which MWI is true and some of them will be ones in which MWI is not true.

There is nothing special about any particular sequence of 0s and/or 1s that could result: any sequence should be represented as much as any other in the set of situations in the reference class corresponding to MWI being true: As there are $2^{1,000}$ possible sequences of 0s and/or 1s, 1 in $2^{1,000}$ of the situations in which MWI is true should correspond to any particular sequence. Similarly, any particular sequence of 0s and/or 1s should be represented as much as any other in the set of situations in the reference class corresponding to MWI *not* being true: As there are $2^{1,000}$ possible sequences of 0s and/or 1s, 1 in $2^{1,000}$ of the situations in which MWI is *not* true should correspond to any particular sequence. Furthermore, it has already been shown that the proportion of MWI and non-MWI situations in the reference class does not change, and as we are dealing with the same proportion (1 in $2^{1,000}$ in this case) of each of *these* proportions, if we just consider the proportion of situations in the reference class, after the sequence of experiments and before you open the cupboard, corresponding to some particular sequence, the proportion of the situations being considered in which MWI is true will be the same as the proportion of situations in the reference class before the sequence of experiments was performed in which MWI is true, and the proportion of situations being considered in which MWI is not true will be the same as the proportion of situations in the reference class before the sequence of experiments was performed in which MWI is not true.

Suppose you now unlock the cupboard, open it and take the printout from inside. You read the particular sequence of 0s and/or 1s on the printout. It starts off something like 01100101011011011011… and goes on, for all 1,000 experiments, with a bit for each. You now know the sequence of outcomes that occurred in all 1,000 experiments. Although this particular sequence was very unlikely, you now know that it occurred. There is nothing mysterious about this, of course: one sequence *had* to occur.

Now that you know the sequence of outcomes that actually occurred in the 1,000 experiments, you know that any situation in your reference class in which this sequence did not occur is no longer a candidate for your situation, so you can remove any such situation from your reference class. This will result in the removal of almost all situations from the reference class: only 1 in $2^{1,000}$ of them will remain. All of the situations that remain in your reference class will now be ones in which the specific sequence on the printout has occurred. A proportion of these will be situations in which MWI is true and a proportion will be situations in which MWI is not true. As has already been shown, the proportion of each kind of situation in the reference class will be whatever it was before the sequence of experiments was started, and in practice this will depend on whatever probability you initially estimated for MWI being true. The experiment has not told you anything.

Suppose now that when you open the cupboard and take the printout from it, you see an unusual sequence. Instead of a mixture of 0s and 1s, the printout contains nothing but 1s. It just goes 11111111111111111111… etc. for all 1,000 experiments. You may suspect some kind of trickery here, but suppose that can somehow be ruled out, so that you just have to accept that this sequence did, in fact, happen. Nothing is different about how you should handle this. You can remove all the situations from your reference class in which anything other than 11111111111111111111… - a 1 in every experiment – occurred. You are now left with a reference class consisting only of situations in which a 1 occurred in every experiment. As before, a proportion of these will be situations in which MWI is true, and a proportion will be ones in which MWI is not true. Again, as has already been shown, the proportion of each kind of situation in the reference class will be whatever it was before the sequence of experiments was started, and what these are will depend on whatever probability you initially estimated for MWI being true. The experiment has told you nothing.

Suppose now that someone walks into the room and tells you that the experimental setup had been tampered with without your knowledge. An automatically triggered execution device had been added to the apparatus. The device consisted of a bomb in the room. If, in any of the 1,000 experiments just performed, the outcome had been 0, a command would have been sent to the bomb, detonating it and killing you. To survive an experiment in the sequence, an outcome of 1 was needed. If you had known this in advance you would have known that the printout in the cupboard *must* have had 1,000 1s on it, because otherwise you would be dead.

If you believe that surviving quantum suicide experiments suggests that MWI is correct, you should now think that this is strong evidence that MWI is correct: you have just survived 1,000 quantum suicide experiments, in each of which you should have only had a 0.5 probability of survival. Your initial probability of MWI being true should be increased as suggested by the Bayesian calculation and you should be left with a situation in which MWI is almost certainly true.

Does this make any sense, though? The fact that you know you are alive means that any of the situations in which 0s occurred in the sequence of experiments can be removed from your reference class, but you have already done this when you opened the cupboard, obtained the printout and saw that the sequence on it was 11111111111111111111… – that the outcome of every experiment had been 1. The only situations now left in your reference class are ones in which every experiment had an outcome of 1. The fact that the bomb was set to explode when a 0 occurred should be irrelevant to this reference class, because a 0 has not occurred in any of the situations in it. All you have found out is that you would be dead now if you were in a situation which is not part of your reference class of possible situations.

The reference class that you now have contains only situations in which a 1 occurred in every experiment. It will still be the case that a proportion of these will be ones in which MWI is true and a proportion will be ones in which MWI is not true, and it has already been shown that the proportion in each case will be what it was from before you started the experiments: your idea of what this should be should depend on your initial assessment of the probability that MWI is correct. The only way that your knowledge of the bomb's existence could change your assessment of the probability that MWI is true is if it alters the proportions of MWI and non-MWI situations in this reference class, but how is it supposed to do that? To do that, the bomb's interaction with the reference class would have to be different for MWI situations and non-MWI situations, yet in both kinds of situation, in the reference class as it now stands, all the experiments have had outcomes of 1 and the bomb has been sitting there not doing anything: the behaviour of the bomb is the same in all situations in the reference class. The fact that the bomb has exploded in situations that are *not* part of the reference class cannot reasonably be expected to affect the composition of the reference class: the reference class merely contains all the possible situations in which 1s occurred in all of the experiments, and to have any effect on the composition of this reference class it would have to affect the way in which you can "put together" a situation in which the bomb did not go off.

Of course, if we consider the situations in a very fine-grained way, the bomb, just by being there and not going off, has an effect on each situation in the reference class, because it is part of the description of the situation, but to think this was of any importance at all would be as absurd as thinking that a houseplant being brought into the room before the experiments started should somehow affect the reference class in any important way. There is no reason to think that any effect of the inactive bomb's

existence in the scenario is biased in any way with regard to MWI or non-MWI situations.

The situation here can also be viewed in terms of *specificity*: situations are represented in the reference class according to their specificity –indicated by the number of ways in which they can occur – and there is no difference in this respect between situations in which MWI is true and you have survived and MWI is false and you have survived.

We can also view this in terms of the *measure reduction* that occurs before you know the result of an experiment being symmetrical. The measure for any particular sequence being obtained is reduced as each experiment happens and the possible situations become more specific: as each experiment is performed, each possible situation that you could have been in just before the experiment gives rise to more specific situations, with slightly longer sequences of outcomes. This may lead us to think that the measure for situations in which MWI is not true becomes progressively smaller, but the problem here is that the same happens for situations in which MWI is true: as each experiment occurs, each situation gives rise to new situations with more information, which are more specific and have less measure. The measure reductions for the MWI and non-MWI situations "keep up with each other", so that you learn nothing from the experiments. Some readers may object that we do know the result in quantum suicide experiments, by still being alive – but that is not being contested: the point has been made that adding the suicide mechanism does not change the resulting reference class. Really, the measure reduction is not the fundamental reason for quantum suicide experiments not telling you anything: the fundamental reason is that the reference class after you have survived and know the result is no different due to the suicide mechanism being in the experiment, but considering the measure reduction for the MWI situations when the result is not yet known gives an idea of what is going on and how the introduction of suicide is not going to alter the final composition of the reference class.

Some readers may object that the introduction of death into this kind of scenario changes things, as you cannot be uncertain of the outcome when one outcome means your death. Such an objection will not work, however: when examined properly, in terms of the reference class of possible situations in which you could be with quantum suicide experiments in your past, you should expect the proportion of MWI and non-MWI situations to be whatever it was before the experiment. We could also answer such an objection by saying that there will always be some delay between the quantum event and death (if it is to happen) occurring, even if this delay is small, so that there must always be a period of time during which the outcome has occurred and the result is unknown. A further answer could be that, before the experiment, you could consider a reference class of your possible situations, and in each situation you could consider information about your future situation to be included: the introduction of suicide into the experiment will not favour MWI or non-MWI in particular. Such answers are not

really needed, however – though they may help some people deal with what may seem to be a counter-intuitive aspect of all this.

5 But what about the Bayesian calculation?

5.1 The Need to Deal with the Bayesian Calculation

The argument in the previous article used a reference class of before and after observer moments. In this article, a stronger argument that surviving quantum suicide would tell you nothing about whether or not MWI is true has been presented in two ways: one way involving a discussion of the "ultimate reference class" of you-centred possible world descriptions, and another way using a less formal idea of possible situations, in which a general idea of specificity based on the number of ways I which a situation can arise is used. The argument that has just been presented is stronger than the one in the previous article because it uses a reference class with much less scope for controversy: the set of all possible situations in which you could be right now. The earlier argument really relies on the same argument that is being used in this article, when it is viewed in terms of specificity increase or measure reduction: a reference class consideration that as worlds split in MWI, before you know the outcome, situations become more specific and assume lower measure, just as they do when MWI is not true – and it is the lack of specificity or measure of a situation that determines its plausibility.

Against this is the Bayesian argument used to justify the idea that surviving repeated quantum suicide suggests that MWI is true. The Bayesian argument involves a calculation that is supposed to show that, by applying the standard Bayesian method, surviving a quantum suicide experiment should cause you to adjust your estimate of the probability that MWI is true upwards – and surviving enough quantum suicide experiments should increase the probability that MWI is true so that it is close to 1 as you want. This would mean that you could use repeated quantum suicide to prove to yourself, for all practical purposes, that MWI is true.

It may seem at this stage that there are two conflicting views of what happens in quantum suicide experiments – the reference class view in the arguments given by the author against quantum suicide telling you anything about MWI and the view in the Bayesian argument for quantum suicide telling you something about MWI – and that both views have some merit. The issue of which view to choose may seem to be a matter of preference – maybe something about how you look at reality – or it may seem that the Bayesian case, because of its reliance on well-established, non-controversial and not very complicated mathematics, is stronger.

For the case made in this article and the previous one to be complete – for the issue to be resolved – the Bayesian calculation needs to be dealt with and shown to be wrong. This is what will be done now. With this issue resolved – with the Bayesian calculation taken out of the game – the arguments that have been made here against the idea that surviving quantum suicide experiments suggests that MWI is true will have no plausible

opposition, and we should think that surviving quantum suicide does not suggest that MWI is true.

5.2 The Bayesian Calculation

Here is a brief description of how the Bayesian calculation is *supposed* to show that surviving quantum suicide suggests that MWI is true.

Suppose you estimate some probability that the hypothesis that MWI is true is correct: this will also mean that you have some probability that the hypothesis that MWI is *not* true is correct. These probabilities are generally known as *Bayesian priors*, and represent your knowledge of the situation before doing any experiments or making any observations.

You perform a sequence of quantum suicide experiments, each of which involves a quantum event with two equally likely outcomes – 0 or 1. If the outcome is 0 in any experiment, you are immediately terminated, but if the outcome is 1 you survive.

Suppose MWI is not true. Your probability of surviving through a long sequence of experiments, each with only a 0.5 probability of survival, is remote – and you can always make this probability as small as you like by arranging to perform more experiments. For example, if you are going to perform 100 quantum suicide experiments, we can say:

$P(\text{Survival}) = 0.5^{100} = 7.9 \times 10^{-31}$.

If the hypothesis that MWI is false is correct, this is the probability that you will observe a future outcome in which you have survived all 100 experiments.

Suppose now that MWI is true. In each experiment, there will always be a future in which you survive and a future in which you do not survive. You will never be aware of being in the future in which you do not survive: it is effectively a future in which you do not exist as an observer, and it can therefore be statistically discounted. The only future in which you can make observations is the one in which you have survived, and therefore, you should always expect to observe an outcome in which you survived. Therefore, for 100 quantum suicide experiments, or any number we want, we can say (according to advocates of this idea):

$P(\text{Survival}) = 1$.

If the hypothesis that MWI is true is correct, this is the probability that you will observe a future outcome in which you have survived 100 experiments, or any number of experiments.

If you really believe that P(Survival) = 1, it literally means that you should not be scared of dying in quantum suicide experiments. The futures in which you do die are ones that you never observe, so you can simply ignore them for statistical purposes. This, of

course, is a controversial idea, and there are some criticisms of it that can be made, but the argument against the Bayesian calculation, as it is performed here, does not rely on such criticisms. It will be shown that, even if we assume for the sake of argument that the P(Survival) = 1 claim for MWI being true is correct, the Bayesian calculation, as it is done here, is actually a misuse of the Bayesian method.

5.3 The general Bayesian method can be reconciled with the reference class view.

The first step in arguing against the use of the Bayesian calculation in this situation is to show that there is no fundamental disagreement between the general reference class approach used here and the general Bayesian method of assigning prior probabilities of some outcome for different hypotheses, and then adjusting them when that outcome occurs.

Suppose you have two hypotheses, Hypothesis 1 and Hypothesis 2, one of which must be true and one of which must be false, and you have estimated prior probabilities, P_P(Hypothesis 1 is true) and P_P(Hypothesis 2 is true) for each being true. The "$_P$" after the "P" is just there to indicate that this is a *prior* probability, assigned before any of the calculations that are about to be performed.

This can be considered in reference class terms. You should think that the proportions of the reference class of possible situations, which you are in right now, corresponding to Hypothesis 1 being true and Hypothesis 2 being true should be given by P_P(Hypothesis 1 is true) and P_P(Hypothesis 2 is true) respectively.

You are about to observe an outcome in the future. The outcome will be one of a number of different possible outcomes: Outcome 1, Outcome 2, Outcome 3, etc. Outcome 1 will be of most interest to us.

If Hypothesis 1 is true, the probability of Outcome 1 occurring is:

P(Hypothesis 1 | Outcome 1)

If Hypothesis 2 is true then the probability of Outcome 1 occurring is:

P(Hypothesis 2 | Outcome 1)

Suppose now that the outcome has occurred, but you do not know which outcome out of Outcome 1, Outcome 2, Outcome 3, etc. it was.

Your reference class is one of situations before the outcome has occurred, but you know that you are in a situation in which one of the possible outcomes has occurred, so your reference class needs updating appropriately. Again, this can be viewed in reference class terms.

Suppose you consider just those situations in the reference class which Hypothesis 1 is true and one of the outcomes Outcome 1, Outcome 2, Outcome 3, etc. has occurred. The proportion of those possible situations in which Outcome 1 has occurred is given by P(Hypothesis 1 | Outcome 1): the probability of Outcome 1 occurring if Hypothesis 1 is true. This is all that the probability means: the probability of Outcome 1 occurring, given that Hypothesis 1 is true, is merely the proportion of "Hypothesis 1 is true" situations in which Outcome 1 has occurred. However, before the outcome had occurred, the proportion of situations in the entire reference class in which Hypothesis 1 is true was estimated to be P_P(Hypothesis 1 is true) and this should not have changed. All this means that the proportion of situations in the entire reference class in which Hypothesis 1 is true was originally P_P(Hypothesis is true), each of these situations has now given rise to a number of new situations in which Hypothesis 1 is true and one of the outcomes Outcome 1, Outcome 2, Outcome 3, etc. has occurred, and the proportion of those situations given rise to by one of the original Hypothesis 1 situations in which Outcome 1 has occurred is P(Hypothesis 1 | Outcome 1). We know, then, that the proportion of all situations in the reference class in which Hypothesis 1 is true and *any* outcome has occurred is P_P(Hypothesis 1 is true) and the proportion of *those* situations in which Outcome 1 has occurred is P(Hypothesis 1 | Outcome 1). Therefore, you can say that the proportion of situations in the entire reference class in which both Hypothesis 1 is true and Outcome 1 has occurred is:

P_P(Hypothesis 1 is true) x P(Hypothesis 1 | Outcome 1)

(Reminder: P_P(Hypothesis 1 is true) is the *prior* probability that Hypothesis 1 is true: the probability that you estimated at the start.)

and a similar argument can be used to show that the proportion of situations in the entire reference class in which both Hypothesis *2* is true and Outcome 1 has occurred is:

P_P(Hypothesis 2 is true) x P(Hypothesis 2 | Outcome 1)

and it has been said earlier that either Hypothesis 1 is true or Hypothesis 2 is true, so the rest of the reference class must be made up of combinations of Hypothesis 1 or Hypothesis 2 and one of the other outcomes, e.g. Hypothesis 1 is true and Outcome 2 has occurred, Hypothesis 1 is true and Outcome 3 has occurred, Hypothesis 2 is true and Outcome 2 has occurred, etc.

So far, although the outcome has occurred, you have not known what it was. It might have been Outcome 1, but it could have been one of the other outcomes – Outcome 2, Outcome 3, Outcome 4, etc. Suppose that you now check to see what the outcome was: it turns out that it was Outcome 1.

The proportion of situations in the entire reference class in which Hypothesis 1 is true and Outcome 1 has occurred is:

P_P(Hypothesis 1 is true) x P(Hypothesis 1 | Outcome 1)

and the proportion of situations in the entire reference class in which Hypothesis *2* is true and Outcome 1 has occurred is:

P_P(Hypothesis 2 is true) x P(Hypothesis 2 | Outcome 1)

and as either Hypothesis 1 or Hypothesis 2 must be true, this covers all of the situations in which Outcome 1 has occurred, so the proportion of the reference class in which Outcome 1 has occurred is given by adding together the two proportions above, and is:

$$\begin{aligned} &P_P(\text{Hypothesis 1 is true}) \times P(\text{Hypothesis 1} \mid \text{Outcome 1}) \\ +\ &P_P(\text{Hypothesis 2 is true}) \times P(\text{Hypothesis 2} \mid \text{Outcome 1}) \end{aligned}$$

You now know that you are in a situation in which Outcome 1 has occurred. As either Hypothesis 1 or Hypothesis 2 is true, you can only be in either one of the situations in which Hypothesis 1 is true and Outcome 1 has occurred or one of the situations in which Hypothesis 2 is true and Outcome 1 has occurred. You can remove all of the other situations from the reference class, e.g. all the situations in which Hypothesis 1 is true and Outcome 2 has occurred, Hypothesis 1 is true and Outcome 3 has occurred, Hypothesis 2 is true and Outcome 2 has occurred, etc. The only situations remaining in the reference class are now those in which Hypothesis 1 is true and Outcome 1 has occurred or Hypothesis 2 is true and Outcome 1 has occurred.

The proportion of situations in the reference class in which Hypothesis 1 is true is now:

$$\frac{P_P(\text{Hypothesis 1 is true}) \times P(\text{Hypothesis 1} \mid \text{Outcome 1})}{\begin{aligned} &P_P(\text{Hypothesis 1 is true}) \times P(\text{Hypothesis 1} \mid \text{Outcome 1}) \\ +\ &P_P(\text{Hypothesis 2 is true}) \times P(\text{Hypothesis 2} \mid \text{Outcome 1}) \end{aligned}}$$

and the proportion of the reference class of your possible situations which is occupied by a kind of situation is the probability that you are in that kind of situation, so:

$$P(\text{Hypothesis 1 is true}) = \frac{P_P(\text{Hypothesis 1 is true}) \times P(\text{Hypothesis 1} \mid \text{Outcome 1})}{\begin{aligned} &P_P(\text{Hypothesis 1 is true}) \times P(\text{Hypothesis 1} \mid \text{Outcome 1}) \\ +\ &P_P(\text{Hypothesis 2 is true}) \times P(\text{Hypothesis 2} \mid \text{Outcome 1}) \end{aligned}}$$

The result of all this is a new value for the probability that Hypothesis 1 is true, P(Hypothesis 1 is true), which may be considerably different to the prior probability, P_P(Hypothesis 1 is true), that was assigned at the start, so based on your observation that Outcome 1 occurred, you have adjusted the probability that Hypothesis 1 is true.

The probability that Hypothesis 1 is true will get reduced considerably if the probability of Outcome 1 occurring, if Hypothesis 1 is correct, P(Hypothesis 1 | Outcome 1), was

very low, while the probability of Outcome 2 occurring, if Hypothesis 1 is correct, P(Hypothesis 2 | Outcome 1), was very high.

The probability that Hypothesis 1 is true will get increased considerably if the probability of Outcome 1 occurring, if Hypothesis 1 is correct, P(Hypothesis 1 | Outcome 1), was very high, while the probability of Outcome 2 occurring, if Hypothesis 1 is correct, P(Hypothesis 2 | Outcome 1), was very low.

The above formula is the one used in the Bayesian calculation used to justify the idea that surviving quantum suicide suggests that MWI is true, but it has been obtained by using the reference class view used in the argument made in this article. In fact, it explains why the Bayesian approach works: the Bayesian calculation is nothing more than a rule based on the reference class reasoning that has just been given. This is important because it suggests that it would be wrong to think that there are two competing views: the reference class view used in the argument being made in this article, which says that surviving quantum suicide tells you nothing about the likelihood that MWI is true, and the view that the situation should be dealt with by the Bayesian calculation, which says that surviving quantum suicide tells you that MWI is more likely to be true. Instead, it should be apparent that the two kinds of view will agree with each other in most situations. Furthermore, the reference class view is the *explanation* for the method used in the Bayesian calculation, so it can hardly be discarded in favour of the Bayesian calculation.

5.4 An Example Bayesian Calculation with a Probability Increase

Let us consider an example of the situation just discussed, involving the two hypotheses, Hypothesis 1 and Hypothesis 2, one of which is true and one of which is false, in which the probability that Hypothesis 1 is true increases considerably.

Suppose you estimate the following prior probabilities for Hypothesis 1 and Hypothesis 2:

P_P(Hypothesis 1 is true) = 0.1
P_P(Hypothesis 2 is true) = 0.9 (This must be 1 – P_P(Hypothesis 1 is true).)

and suppose that, given that Hypothesis 1 is true, the probability of Outcome 1 occurring is as follows:

P(Hypothesis 1 | Outcome 1) = 0.999

and suppose that, given that Hypothesis 2 is true, the probability of Outcome 2 occurring is as follows:

P(Hypothesis 2 | Outcome 2) = 0.001

so if Hypothesis 1 is true we expect Outcome 1 probably to occur, almost certainly, but if Hypothesis 2 is true we expect Outcome 1 probably not to occur.

Suppose now that you observe that Outcome 1 has occurred. You can now calculate an adjusted probability for Hypothesis 1 being true as previously described.

$$P(\text{Hypothesis 1 is true}) = \frac{P_P(\text{Hypothesis 1 is true}) \times P(\text{Hypothesis 1} \mid \text{Outcome 1})}{P_P(\text{Hypothesis 1 is true}) \times P(\text{Hypothesis 1} \mid \text{Outcome 1}) + P_P(\text{Hypothesis 2 is true}) \times P(\text{Hypothesis 2} \mid \text{Outcome 1})}$$

$$= \frac{0.1 \times 0.999}{(0.1 \times 0.999) + (0.9 \times 0.001)}$$

$$P(\text{Hypothesis 1 is true}) = 0.991$$

Your observation that Outcome 1 has occurred has changed things completely: it has changed the situation from one in which it is unlikely that Hypothesis 1 is true (a probability of 0.1) to one in which it is very likely that Hypothesis 1 is true (probability of 0.991).

5.4.1 An Example of the Bayesian Calculation with Quantum Suicide

Here is an example of the situation just discussed which specifically involves quantum suicide. Let us say that Hypothesis 1 = MWI is true and Hypothesis 2 = MWI is not true.

Suppose you estimate the following prior probabilities:

$P_P(\text{MWI is true}) = 0.02$
$P_P(\text{MWI is not true}) = 0.98$

i.e. you think there is a 0.02 probability of MWI being true: you accept the possibility, but you do not take it very seriously.

Suppose that you are going to perform 100 quantum suicide experiments, each of which should involve a 0.5 probability of survival. You think that if MWI is true you should be sure of observing that you are alive in the future, so:

$P(\text{MWI is true} \mid \text{Survival}) = 1$

But if MWI is *not* true, you expect the probability of observing that you are alive in the future to be somewhat lower:

$P(\text{MWI is not true} \mid \text{Survival}) = 0.5^{100} = 7.9 \times 10^{-31}$

Suppose now that you find yourself alive after the 100 quantum suicide experiments. You can now calculate an adjusted probability for MWI being true as previously described.

$$P(\text{MWI is true}) = \frac{P_P(\text{MWI is true}) \times P(\text{MWI is true} \mid \text{Survival})}{P_P(\text{MWI is true}) \times P(\text{MWI is true} \mid \text{Survival}) + P_P(\text{MWI is not true}) \times P(\text{MWI is not true} \mid \text{Survival})}$$

$$= \frac{0.02 \times 1}{(0.02 \times 1) + (0.98 \times 7.9 \times 10^{-31})}$$

$$P(\text{MWI is true}) = 0.99999999999999999999999999996$$

(A large number of significant figures were just provided here to show how close the value is to 1, without giving the idea that it is exactly 1.)

Your observation that you have survived has changed things completely: it has changed the situation from one in which it is unlikely that MWI is true (a probability of 0.02) to one in which it is almost certain that MWI is true (a probability of 0.99999999999999999999999999996).

But is this correct? We might disagree with the idea that P(MWI is true | Survival) = 1, but the argument against use of the Bayesian approach in quantum suicide will not depend on such disagreement. We need an argument that can be accepted without controversy, so it should not rely on anyone's particular ideas about continuity and survival. We need to look at the "mechanism" by which a probability can be increased when an outcome is observed.

5.5 The "Mechanism" By Which a Prior Probability Can Be Increased On Observation of an Outcome

The justification for the Bayesian calculation has been given, and it has been shown that observation of an outcome can justify an increase in the probability that a hypothesis is true, but we need an understanding of the basic "mechanism" that can cause the probability of a hypothesis being true to increase when a particular outcome is observed: we need to check that this "mechanism" is still there in quantum suicide experiments.

Let us consider the situation previously discussed in 5.3, in which there are just two hypotheses – Hypothesis 1 and Hypothesis 2 – one of which must be true, and one of a number of outcomes – Outcome 1, Outcome 2, Outcome 3, etc. is about to occur.

The reference class initially consists of situations in which Hypothesis 1 is true and situations in which Hypothesis 2 is true, and you should expect the proportion of each to

correspond with the prior probabilities, P_P(Hypothesis 1 is true) and P_P(Hypothesis 2 is true), that you assign.

When you know that the outcome has occurred your reference class consists of situations in which the outcome has occurred: each situation in the reference class corresponds to either Hypothesis 1 or Hypothesis 2 being true and one of the outcomes, Outcome 1, Outcome 2, Outcome 3, etc.

Suppose that the probability of Outcome 1 occurring, given that Hypothesis 1 is true, is a very high value and the probability of Outcome 1 occurring, given that Hypothesis 2 is true, is a very low value. That is to say:

P(Hypothesis 1 | Outcome 1) = some very high value
P(Hypothesis 2 | Outcome 1) = some very low value

These probabilities mean that a very high proportion of the situations in the reference class in which Hypothesis 1 is true will be ones in which Outcome 1 has occurred, while a very low proportion of the situations in the reference class in which Hypothesis 2 is true will be ones in which Outcome 1 has occurred.

Suppose now that you have found out that Outcome 1 has occurred. You can remove any situations from your reference class in which Outcome 1 has *not* occurred, leaving only those situations in which Outcome 1 has occurred. Let us consider what happens during that removal process.

Let us first consider the removal of situations in which Hypothesis 1 is true.

Of all the situations in which Hypothesis 1 is true, only those in which Outcome 1 has occurred will remain in the new reference class: the rest will get removed. The proportion of those that remain is therefore the proportion of these situations in which Outcome 1 has occurred, and this is given by P(Hypothesis 1 | Outcome 1). This is a very high value, so most of the situations in which Hypothesis 1 is true are ones in which Outcome 1 is true and which will remain in the new reference class.

Let us now consider the removal of situations in which Hypothesis 2 is true.

Of all the situations in which Hypothesis 2 is true, only those in which Outcome 1 has occurred will remain in the new reference class: the rest will get removed. The proportion of those that remain is therefore the proportion of these situations in which Outcome 1 has occurred, and this is given by P(Hypothesis 2 | Outcome 1). This is a very *low* value, so few of the situations in which Hypothesis 2 is true are ones in which Outcome 1 is true and which will remain in the new reference class.

The end result of this is a new reference class which is made by combining all of the Hypothesis 1 situations which are to remain and all the Hypothesis 2 situations which are to remain. A large proportion of the Hypothesis 1 situations have remained (those in

which Outcome 1 has occurred), while only a small proportion of the Hypothesis 2 situations have remained (again, those in which Outcome 1 has occurred). The effect of this will be to increase the proportion of Hypothesis 1 situations in the reference class. The probability that Hypothesis 1 is true will therefore have increased. This might seem to suggest that anything that increases P(Hypothesis 1 | Outcome 1) must always cause P(Hypothesis 1 is true) to be higher when Outcome 1 is observed, but we need to be careful here, as will be shown.

The indication that a large proportion of the Hypothesis 1 situations were going to remain in the reference class was the fact that P(Hypothesis 1 | Outcome 1) was a high value, but the probability is measuring the number of situations in which Hypothesis 1 is true and Outcome 1 has occurred *as a proportion of all the situations in which Hypothesis 1 is true.* What is really important, however, is how many of these situations are going to end up in the new reference class with the Hypothesis 2 situations. The probability is measuring one part of the "Hypothesis 1 is true" part of the reference class (the part corresponding to situations in which Outcome 1 has occurred) against all of that part of the reference class in which Hypothesis 1 is true, but that is only relevant because it should give an indication of how many of the Hypothesis 1 situations are going to end up in the new reference class with the Hypothesis 2 ones as a proportion of all the Hypothesis 1 and Hypothesis 2 situations that end up in the new reference class. What matters is the *specificity* of the situations in which Hypothesis 1 is true and Outcome 1 has occurred, relative to the specificity of all of the situations in which Outcome 1 has occurred: a low specificity implies a high probability.

That may seem a bit involved, but it becomes easier to consider if we think of it in terms of *numbers* of situations. A certain number of situations in the reference class are ones in which Hypothesis 1 is true. Given that Hypothesis 1 is true, the probability that Outcome 1 has occurred is very high, so this means that a large number of the Hypothesis 1 situations are ones in which Outcome 1 has occurred, so a large number of these situations are going to remain in the new reference class. The probability P(Hypothesis 1 | Outcome 1) is important only in that indicates that a large number of Hypothesis 1 situations are going to survive.

It may seem that the probability that Outcome 1 has occurred, given that Hypothesis 1 is true, P(Hypothesis 1 | Outcome 1), will always correspond with the idea of the proportion of Hypothesis 1 situations in which Outcome 1 is true as a proportion of all Outcome 1 situations and with the idea of the absolute number of Hypothesis 1 situations that are Outcome 1 situations, and which therefore survive into the new reference class. P(Hypothesis 1 | Outcome 1) should be an indication of the *overall* specificity of Hypothesis 1 and Outcome 1 situations, and it generally will be. Usually, anything that increases the probability that, given that Hypothesis 1 is true and Outcome 1 has occurred, will increase the *overall* representation of situations in which Hypothesis 1 is true and Outcome 1 has occurred and in particular, critically, will increase the representation of situations in which Hypothesis 1 is true and Outcome 1

has occurred in the part of the reference class corresponding to Outcome 1 occurring for Hypothesis 1 or Hypothesis 2 – the part of the reference class that remains when the new reference class is made.

As an example, suppose there are two drugs, Drug A and Drug B, each of which has a small probability of curing a patient of some disease. One of the drugs is selected by tossing a coin and administered to a patient. The patient is cured. We would now have a reference class made up of situations in which Drug A was used and the patient was cured and Drug B was used and the patient was cured.

Suppose we repeat the process, but we increase the effectiveness of Drug A, so that it is much more likely to cure a patient. We then toss a coin to select a drug and treat a patient with the selected drug, again. As before, the patient is cured. Before we know that the patient is cured, a higher proportion of the entire reference class of possible situations correspond to situations in which the patient was given Drug A and cured: the proportion of situations in which the patient was given Drug A has stayed the same (0.5 from the coin toss) and out of these the proportion in which the patient was cured was higher. There is therefore a correspondence between the probability that Drug A cures the patient and the representation of situations in which the patient is cured by Drug A in the reference class relative to the representation of situations in which the patient was cured by Drug B – and this is important because, when we find out that the patient was cured, it is these two kinds of situations that combined to make the new reference class. The increase in the probability of Drug A curing a patient is therefore adding situations in which the patient is cured by Drug A to the reference class. It is increasing the number of ways a patient can be cured by Drug A, and decreasing the specificity of the kind of situation in which a patient is cured by Drug A. This is what normally happens when we increase probabilities: increasing probabilities tends to mean adding situations to reference classes. It is why the Bayesian approach generally works.

There is a problem with quantum suicide, however, that results in a disconnect between the P(Survival) = 1 probability and any real low specificity in the reference class and there are three ways of viewing it.

5.6 Quantum suicide involves a disconnect between P(Survival) = 1 and any *real* low specificity in the reference class: three views.

5.6.1 A Summary of the Three Views

Viewed the first way, the problem with trying to use the Bayesian approach with quantum suicide relates to the actual measures taken in setting up the experiment to result in P(Survival) = 1: it can be shown that they are non-standard because they involve taking outcomes out of the reference class that are irrelevant to the probability calculated after the outcome. Viewed the second way, the problem also relates to the

actual measures taken in setting up the experiment to result in P(Survival) = 1: they are non-standard because they have no special effect on the *histories* of the situations which they are supposed to be making more probable, and can therefore not affect the specificity of these situations. Viewed the third way, the problem relates to the way in which the probability P(Survival) = 1 is assigned: this too can be shown to be non-standard. All three of these are the same problem – just looked at from a different perspective. Each will now be discussed.

5.6.2 View 1: With quantum suicide, making one of the outcomes more likely is being done in a non-standard way that merely removes irrelevant situations from the reference class.

To see why this is the case, let us first consider an experiment which is somewhat like a quantum suicide experiment, except the device to kill you is disabled, so there will not be any actual suicide. Let us imagine a scenario like the one described earlier in 4.4. You are in a room with a bomb. A series of experiments are run, in each of which a quantum event has an outcome of 0 or 1. There is a mechanism that causes the bomb to detonate if an outcome of 0 occurs, but for the present the bomb is disarmed, so that you are not at risk of death. If the outcome is 1, nothing happens (and would not happen even if the bomb were armed). As before, the outcomes of the experiments are printed out by a printer inside a cupboard. 100 experiments are to be performed, each generating an outcome of 0 or 1. If the bomb were armed, the only way you would survive the sequence of experiments would be if every experiment had an outcome of 1, so the outcomes would be 1111111111111111111… etc. As the bomb is currently, disarmed, this is not yet an issue. If MWI is true, the probability of a 1 occurring in any particular experiment is 0.5:

P(MWI is true | Outcome = 1) = 0.5

You run the sequence of 100 experiments, but you do not yet look in the cupboard. The printout in the cupboard now contains a sequence of 100 0s and/or 1s, e.g. 10111001010100010110… etc.

Suppose you have assigned some probability to MWI being true. Your reference class will consist of possible situations in which MWI is true and possible situations in which MWI is not true. For each situation in which MWI is true, there will be situations corresponding to every possible sequence of 1s and 0s that could have been printed out. For each situation in which MWI is not true, there will also be situations corresponding to every possible sequence of 1s and 0s that could have been printed out.

Suppose you are going to repeat the sequence of experiments and you have a bet with someone that a 1 will result in every experiment – that the printout will contain a sequence of 100 1s and no 0s – and you are allowed to make some changes to the experimental procedure. One way you could do this would be to alter the apparatus

somehow so that each experiment is actually more likely to produce a 1: you would have to make it very likely that each experiment produces a 1 to have a realistic chance of getting a sequence of 100 1s.

You decide on an unusual method: before starting the sequence of experiments, you arm the bomb, so that if a 0 occurs in any experiment, it will detonate and kill you before you have time to become aware of what has happened. You reason that, if MWI is true, as you will never observe a future in which the bomb detonates – in which the outcome of an experiment was 0 – such futures do not exist for you and the only futures that exist are ones in which the outcome was 1. You can therefore say that if MWI is true, the probability that you will survive any given experiment is 1:

P(MWI is true | Survival) = 1

which is another way of saying that, if MWI is true, the probability that the outcome of any given experiment will be 1 is 1:

P(MWI is true | Outcome = 1) = 1

When MWI is true, the probability that the outcome was 1 has increased from 0.5 to 1. This means that a larger proportion of the situations in the reference class for which MWI is true are now ones in which the outcome was 1. The problem here is that *this proportion is only a proportion of the part of the reference class in which MWI is true: the proportion of all of the situations in the reference class in which the outcome was 1 remains the same*. Normally, altering the first kind of proportion would alter the second, because increasing the proportion of the part of the reference class corresponding to a particular hypothesis being true which is taken up by a particular outcome occurring implies that some of the situations in which that hypothesis is true and that outcome has not occurred have been changed into situations in which that hypothesis is true and that outcome has occurred: this is the only way that you could get the proportion to increase if the proportion of the entire reference class in which the hypothesis is true remains constant. Therefore, situations in which the hypothesis is true and the outcome has occurred have been added to the reference class. There is an assumption here, however: this only applies if the proportion of the entire reference class in which the hypothesis is true remains constant, and that is exactly what does *not* happen when the experimental procedure being considered here is changed by arming the bomb to detonate if an outcome of 0 occurs. If we accept that P(MWI is true | Outcome = 1) = 1, it may seem that some of the situations in the reference class in which MWI is true and the outcome is 0 have been changed into situations in which MWI is true and the outcome is 1, but this would only be the case if the proportion of the reference class in which MWI is true had remained the same size: in fact, the proportion of the reference class in which MWI is true has been reduced by removing those situations in which MWI is true and the outcome is 0 – by ensuring that you have been blown up in those situations. The fact that the proportion of the situations in which MWI is true for which

the outcome was 1 has increased is irrelevant, because this does not mean anything in any other context – outside that part of the reference class. In reality, *the proportion of all situations in which the outcome was 1 for which the outcome was 1 and MWI is true has not altered at all* – and it is this proportion that is important, because it gives the probability that MWI is true. The change from P(MWI is true | Outcome = 1) = 0.5 to P(MWI is true | Outcome = 1) = 1 may seem to have made the kind of situation in which MWI is true and the outcome is 1 less specific, but it has not really done any such thing.

What has happened here is that there is normally a correspondence between the probability that an outcome has occurred given that some hypothesis is true and the proportion of *all* situations in the reference class in which that outcome occurred for which the hypothesis is true, but the unusual step of killing yourself means that there is now a disconnect between the probability and the relevant proportion of the reference class.

As an analogy, suppose a company, Spaceley's Sprockets, sells sprockets, and competes with Cogswell's Cosmic Cogs, that also sells sprockets: sprockets are a generic product. The owner of Spaceley's Sprockets is concerned that his company is not making enough money, in comparison with other companies in general, so he decides to increase his market share. He pays someone to find out the details of people who are about to buy sprockets from Cogswell's Cosmic Cogs, and then contact those people and persuade them they do not want sprockets at all by telling them how useless sprockets are or telling them horror stories about some of the (alleged) ill-effects of sprockets on people around them.

This seems to be successful. A huge number of people who were about to buy sprockets from Cogswell's Cosmic Cogs are persuaded to change their minds. The owner of Spaceley's Sprockets sees that his company's market share has increased, and this seems to suggest that his company is doing much better. This, of course, is an illusion: if he looked at the number of sprockets his company was selling, he would see that there has been no change. The company's market share was only increased by reducing the size of the market – by removing sprocket purchases from other companies from the market. Spaceley's Sprockets is doing better, in a sense, but only in the limited context of the specific market in which it operates. In the wider context of business and the economy in general – in comparison with other companies – the company is not doing any better. The owner of Spaceley's Sprockets increased his market share in a non-standard way – one that was purely destructive.

5.6.3 View 2: The measures taken in setting up the experiment to result in P(Survival) = 1 are non-standard because they have no special effect on the history of the situations which they are supposed to be making more probable.

When things are done to affect the probability of an outcome, they generally affect the *history* of a situation in which that outcome occurs in such a way a way that there are

more ways for that outcome to occur. As each situation in which the outcome has occurred needs a different history, having more ways available for the situation to occur generally corresponds to more situations in which the outcome has occurred.

Earlier, the example of a drug to deal with a disease was given, with the drug being made more effective so that it has a higher probability of curing a patient. Increasing the effectiveness of a drug would change the history of a situation in which a patient has been cured by the drug: the history would be one in which the patient's body had contained a more effective drug, doing different things inside his body.

As another example, you may have a bag of red and green balls, and you may want to increase the probability that, if you take a ball randomly from the bag, you take a red ball. You could do this by increasing the number of red balls, decreasing the number of green balls or changing some green balls into red balls (which effectively means both increasing the number of red balls and decreasing the number of green balls). All of these would affect the history of a situation in which a red ball is drawn from the bag. For example, if the bag contains a greater proportion of red balls, the hand of someone reaching into the bag will bump against more red balls and fewer green balls, before eventually selecting a ball. The actual physical process involved in taking a red ball from a bag with a big proportion of red balls in it is different.

This is not the case with the method of setting yourself up to be killed in order to ensure that the probability of one of the outcomes is 1 if MWI is true. The mechanism that is set up to kill you is supposed to kill you when the outcome that you do *not* want to experience has occurred, but when the outcome that you *do* want has occurred, this mechanism has *not* intervened in its history.

Suppose the experiment has two possible outcomes, 0 and 1, and you set things up so that a bomb will detonate if the outcome is 0, reasoning that, if MWI is true, P(Outcome = 1) = 1,which means the same as P(Survival) = 1. The only time the bomb will detonate is if the outcome is 0, In a future situation in which the outcome was 0, the bomb will have detonated, but you will not be alive to experience such a future and you have statistically discounted it: this situation, in any event, will not be in the reference class of situations that determines the final probability. In any situation in which the outcome is 1, the bomb will have done nothing. The bomb's role in the history of any situation in which the outcome is 1 and you survive is clearly trivial: it is the same role that could be taken by a houseplant that was placed in the room with you during the experiment. The inactive bomb is not doing anything important to *participate* in the history of the outcomes in which you survive.

Some readers may object that it is only the author's intuitive judgement that the inactive bomb does not participate in any important way in the situations in which you survive. Who is to say that an inactive bomb is not important? It should be fairly obvious that it is not important – that the inactive bomb is not going to have any major

effect on the future situations corresponding to some present situation. A clear sign that the inactive bomb is not important in the history of any situation where it remains inactive is that it is present – inactive – in *all* the situations in which you survive – including the ones for which you are not attempting to increase the probability of survival.

We can consider this in the context of the imaginary scenario with Spaceley's Sprockets and Cogswell's Cosmic Cogs that was used in 5.6.2: The owner of Spaceley's Sprockets is interfering with his competitor's sales to increase his market share, but he is not getting any of these sales himself. Any benefit is illusory. An obvious indication of this is that his plan of getting someone to persuade potential customers of Cogswell's Cosmic Cogs that they do not want sprockets involves events that play no important part in the history of a sprocket purchase from Spaceley's Sprockets.

5.6.4 View 3: With quantum suicide, P(Survival) = 1 involves a non-standard way of assigning probability.

Instead of viewing the situation in terms of actual changes to the experimental procedure to cause P(Survival) = 1, we can view it in terms of how the probability of survival, given MWI, P(MWI is true | Survival) = 1, is assigned: it is actually assigned in a non-standard way.

Suppose that you have the experiment just discussed in 5.6.2, already set up, so that there will be a sequence of 100 experiments, each with an outcome of 0 or 1, and the bomb will be detonated if an outcome of 0 occurs. This time we will not worry about changing the experimental procedure to cause P(MWI is true | Outcome = 1) = 1: the relevant changes have already been made. Instead, we will consider the actual assignment of the probability P(MWI is true | Outcome = 1) = 1 or P(MWI is true | Survival) = 1.

Each experiment has two possible outcomes – 0 or 1 – and the bomb detonates if an outcome of 0 occurs. Looking at this from your perspective before a bomb detonation, there are two possible outcomes: 0 or 1. If MWI is true, you decide that the future in which Outcome = 0 can be statistically discounted, as you are dead in this future and are not in a state that allows you to observe anything. All of the probability that the 0 outcome would have had therefore goes to the 1 outcome. This is a non-standard way of assigning probability, because an outcome is not normally statistically discounted, and gets its probability transferred over to another outcome, on the basis that you are not in such a state that you can observe it. Normally, the probability of an outcome is supposed to give an indication of the specificity of that outcome (high probability corresponding to low specificity): it is supposed to be about the physical characteristics of what has to happen to bring about that outcome, and how specific this is – what proportion of all the different ways that things could happen that it constitutes – or about the physical state that your situation has to be in – what proportion of all the different ways your situation could exist that it constitutes. However, when you give one

outcome a probability of 1 because you are not going to be able to observe another outcome, the outcome getting the probability of 1 is not getting it because of its own lack of specificity: it is getting it because of the state that *you* are in when another outcome occurs. It is all about you – not the outcome that is getting all the probability. It may seem, here, that the author is arguing against P(Survival) = 1, but that is not the case. The author recognizes P(Survival) = 1 as controversial, but is accepting it for the purposes of argument. The issue is that P(Survival) = 1, if it is valid, is about *you*. It is about the state in which *you* are going to be in one of two possible futures, and not about the outcome that gets the probability in one of those futures. Because of this, even if P(Survival) = 1 is valid, the probability has become disconnected from the actual specificity of the outcome itself, but the Bayesian method relies on the assumption that there is a correspondence between the probability of the outcome and its specificity. In fact, whatever probability you assign for P(Survival) is irrelevant to the issue of whether MWI tells you anything about the likelihood that MWI is true: all that matters is the composition of the reference class for a survivor – and probability assignments of any kind before experiments are not going to change this.

6 Some Analogies

The argument that has been given has been based on reference class, and it is that if you consider the reference class of all of your possible situations, and determine the proportion of situations in this reference class that involve living in MWI realities, if you perform quantum suicide experiments, and survive them, the proportion of situations in which MWI is true will remain the same. This is because, although removing the situations in which MWI is true and a particular outcome occurs, by suicide and the probability assignment P(Survival) = 1, increases the representation of the situations in which MWI is true and the other outcome occurs relative to the representation of the situations that were removed, it does not increase the representation in the reference class of the situations in which MWI is true and this outcome occurs relative to the representation of the situations in which MWI is not true and the same outcome occurs. P(Survival) = 1 can only be true, if we accept it as true, because of a disconnect between the two kinds of representation, and between the probability and the second kind of representation.

The suicide component of the experiment is merely removing situations from the reference class that will not be part of the reference class of someone who has observed the outcome corresponding to survival anyway. This can be viewed in terms of specificity: it is the specificity of a situation that determines your probability of being in it, all else being equal, and the outcome of some quantum event in MWI is just as specific as the corresponding outcome in the absence of MWI. Another way of viewing this is in terms of measure reduction: splitting of worlds in MWI, before you know the result of an experiment, involves a decrease in measure, just as the measure decreases for any individual outcome in a probability distribution in the absence of MWI.

Some analogies for this will now be provided. The author does not expect these analogies to be safe from any argument that can be imagined. People may be able to find various issues with them: it might be suggested that they do not exactly match the issue of MWI vs. non-MWI. The purpose of the analogies is not to give an accurate representation of the MWI vs. non-MWI issue with quantum suicide. Rather, it is to give an idea of the sort of thing that is going on here, and to show how situations can occur where we may make an assumption about continuity – survival – that could lead us into Bayesian calculations in which probabilities of outcomes are computed in a way that is disconnected from the specificity of those outcomes, leading to incorrect results – how our intuition might fail us.

6.1 Analogy 1: The Clone Factories

You have been abducted by a secret society that likes, for some reason that is secret, to perform secret experiments on people. They tell you that you are going to be experimented on in a *Type-1* factory or a *Type-2* factory.

In a Type-1 factory, you will be processed with n other abducted people. You will be subjected to 100 "events", one event occurring each week. In each event you will be sedated, a coin will be tossed, and then you will be woken up. Note that the coin does not actually do anything here: it was just put in as a redundant prop. The point is that you are guaranteed to survive each "event". The coin toss cannot kill you. After the sequence of events, you will wake up in a green room, with a sign on the wall saying "You survived."

So, and fairly non-controversially, P(You'll be alive at the end of this) = 1.

We do not even need to worry about what constitutes continuity here, unless being sedated is an issue. You really are just going to sleep and wake up repeatedly. The Type-1 factory is analogous with MWI.

(There is one thing feature makes this a bit unlike MWI quantum suicide: your measure does not decrease in each event, but stays constant. To fix this, we will do something in the Type-2 factory to make it consistent.)

In a Type-2 factory, you will also be processed with n other abducted people. You will be subjected to 100 "events", one event occurring each week. In each event you will be sedated, and a coin will be tossed. If the coin toss result is "heads" a clone (assume it is an exact or near-exact brain copy) will be made, and both the original version of you and the clone will be woken up: you will then both be allowed to live for a week before going through to the next "event". If the coin toss result is "tails" you will be killed, while asleep and without being cloned. "Heads" is a nicer result for you. If you survive through the entire sequence, you will wake up in the green room with the sign on the wall saying "You survived."

(Why the cloning here? In a Type-2 factory, you do not get the reduction in measure that we should expect to occur in MWI. To compensate for this, and to stop it skewing the statistics, if you survive in a Type-2 factory you get an increase in measure to compensate for the decrease on the occasions when you get killed. The overall result is that a Type-1 factory and a Type-2 factory should have the same effect on the measure of observers.)

Here, for each event, P(Survival) = 0.5, and

P(You'll be alive at the end of this) = $0.5^{100} = 7.9 \times 10^{-31}$

Your chances of survival in a Type-2 factory are very low, in contrast with the certainty of your survival in a Type-1 factory. Someone may try to make an argument that the survival of your clone is somehow good enough to constitute continuity for you, so that you should not regard yourself as really being killed if your clone survives, but note that *the coin toss takes place a week after the clone was made*: both of you have been alive and having different experiences for a week before this occurs, which should seem to

give enough divergence that you will regard your own future as being distinct from that of the clone.

In both types of factory, you are kept in solitary confinement. You never see any of the other people, and for a Type-2 factory this includes any clones of you, or anyone who was cloned to make you. The experience is basically the same in a Type-1 factory and a Type-2 factory: you are kept in solitary confinement and repeatedly sedated until you end up in the green room with the sign, unless you are killed first.

Clearly, if you can choose a type of factory in which to be, you should choose a Type-1 factory: you will tend to end up less dead.

Now, suppose you have just woken up in the green room with the sign saying "You survived." Are you in a Type-1 factory or a Type-2 factory? Your calculation of survival probability for the Type-1 factory is of no help here. What you should want to know is: how many people wake up in the green room in a Type-1 factory and how many wake up in the green room in an Type-2 factory? If one Type-1 factory were in use and one Type-2 factory were in use, and n were the same in each case, you would have no way of knowing which factory you were in: the experience of waking up in the green room would be just as common in each case. If n were larger for one type of factory, or one type of factory were in more common use, it would change things.

Suppose your abductors are only going to use one factory for everyone. It is going to be a Type-1 factory or aType-2 factory. They will decide which factory to use by tossing a coin, and the other factory will not process anyone. Here, the nature of each factory changes nothing. Each factory has a 0.5 probability of being chosen as the factory that will be used, and once chosen, will be expected to generate the same number of green room survivors.

Your memory of making the P(you'll be alive at the end of this) = 1 calculation is telling you nothing: the number of people expected to wake up in the green room in each case is the same, and you should think it just as likely that you have awoken in the green room of a Type-2 factory remembering that, unless one factory is more common, more likely, or processes more people. When you wake up in the green room, your estimate of the chances of being in a Type-1 factory or a Type-2 factory should be completely dominated by what you know about the expected number of people waking up in the green room in each case. You should want to know about the number of people processed in each factory, how many factories of each type are in use and, if some chance event has been used to select one factory or the other, what the probabilities were. Ultimately it all comes down to reference class. Ideally, a better reference class than this would be used (and the ultimate reference class, suggested earlier, is a better reference class: in fact you could say that the ultimate reference class of you-centred possible world descriptions is a subset of this reference class anyway, so all else being equal, more observers in the green room should correspond to more possible

candidates for situations completely consistent with yours in the ultimate reference class) but this should make the point. Likewise, your expectations regarding MWI or non-MWI being true if you survive a sequence of quantum suicide experiments should be completely dominated by what you already know about the probability of each view being correct.

6.2 Analogy 2: The AI Laboratories

This analogy has some similarity with the clone factory analogy that was just given in 6.1, but it now involves a proper measure reduction resembling what we should expect in MWI – at the expense of allowing possible objections about computers and minds and whether copies count.

Suppose there are business plans to set up two artificial intelligence (AI) laboratories, only one of which will be implemented. The first laboratory is a *Type-1* AI laboratory. The other laboratory is a *Type-2* AI laboratory.

In both laboratories, AIs will be manufactured. These will be simulations of minds like ours, running in virtual reality simulations.

In the Type-1 AI laboratory, the plan is to make 1,000 *identical* copies of the same AI system, and run each copy in an identical virtual reality. The experiences of each of the 1,000 AIs will therefore be the same. After one week of existence, all the AI systems will be fed into a computer known as an "AI blender" and it will produce one AI system whose experiences pick up where the experiences of each of the copies left off. Before going into AI blender, the AI systems will all know they are destined to go into AI blender. The last memory of each the original 1,000 copies will be when it is just about to be fed into the AI blender. The first experience of the program that comes out of the AI blender will be that of being in a green room with a sign on the wall saying "You survived." It will remember events before being fed into AI blender.

In the Type-2 AI laboratory, the plan is to make 1,000 *different* AI systems, each running in a virtual reality. The experiences of each AI will be different. As before, each AI program knows it is destined to be fed into the AI blender. As before, only one program emerges from AI blender. This is a randomly selected AI from the 1,000 AIs that were fed in to AI blender: the other 999 AIs are simply erased. As before, the first experience of the program that comes out of the AI blender will be that of being in a green room with a sign on the wall saying "You survived," and it will remember events before being fed into AI blender.

Just one AI laboratory – Type-1 or Type-2 – has been selected to be set up. A coin was tossed to make the decision. If the coin toss result were “heads”, then a Type-1 AI laboratory was to be set up. If the coin toss result were “tails”, then a Type-2 AI laboratory was to be set up.

Suppose you are an AI about to be fed into AI blender and you do not know the result of the coin toss: you do not know if the Type-1 AI laboratory or the Type-2 AI laboratory was built, so you do not know which kind of laboratory you are in. You may think that the Type-1 AI laboratory is a nicer place to be. You can be sure that, as all the AIs going into AI blender are identical, the single AI that survives will have your memories: it will have the memories of all 1,000 of your "siblings" as you are all the same. You may therefore think that this means there is a future in which you continue and that P(Survival) = 1. This, of course, is going to be controversial. Some people may say that the copy cannot be a valid continuation of the original. Others may say that there is a loss of measure here, as many identical AIs are reduced to one AI, and that this somehow reduces the probability of continuation. These issues should not concern us here. The point is that we might think that some kind of case can be made for saying that P(Survival) = 1, and we are going to look at how this should relate to what we know in the future.

With the Type-2 AI laboratory, the situation is philosophically less controversial: you are one of 1,000 different AIs, only one of which will correspond to the copy that finds itself in the green room. You may therefore think that P(Survival) = 0.001.

Suppose you find yourself conscious, in the green room with the sign on the wall saying "You survived." If you used the same kind of Bayesian reasoning used to justify the idea that surviving quantum suicide suggests that MWI is more likely to be true, you might think that this is strong evidence that the Type-1 AI Laboratory was the one that was built, and you have just been through its process. Your memory tells you that, before going into the AI blender, P(Survival) = 1 for the Type-1 AI laboratory and only 0.001 for the Type-2 AI laboratory, so it is hard to see how you can seriously think you are in the Type-2 AI laboratory. A naive Bayesian approach says that the probability of the outcome that you are experiencing was larger for the "You are in the Type-1 AI laboratory" hypothesis, and that this supports this hypothesis.

This idea would be completely wrong, and we should be able to see this easily. As has been said, whatever you should expect to happen before going into AI blender could be a matter of controversy. Some people might say that P(Survival) = 0 in both cases: that copies cannot constitute survival and that the blending process is fatal. Others may say that the two situations are not equivalent: that 1,000 identical versions of you is somehow just like one, and that this amounts to a pre-existing bias in favour of the Type-2 AI Laboratory as it is starting with 1,000 versions of you whereas the Type-1 laboratory starts with just one. The author would disagree with this, as will be evident from his previous writing on the subject, but this does not matter here. There is no need to have any argument at all about continuation and survival probabilities: for the purposes of what is being discussed here, it is all irrelevant. We only need to look at the result. In both AI laboratories, exactly one observer ends up in the green room with the sign saying, "You survived." When you find yourself in the green room with the sign, you know you belong to a reference class of observers in that kind of situation, and the

version of you in the Type-1 AI laboratory and the version of you in the Type-2 AI laboratory are equally represented in this reference class. There is no way that finding yourself in the green room with the sign is going to distinguish between the two hypotheses, here, when, clearly, the existence of one observer in such a situation occurs in both cases. You might remember thinking "Hang on! I computed that in the Type-1 AI laboratory P(Survival) = 1," but regardless of whether the Type-1 AI laboratory or the Type-2 AI laboratory was set up, there will be exactly one observer in that room having that memory. P(Survival) = 1 is deceptive, because it would normally suggest that the Type-1 AI laboratory has more green rooms, with more AIs finding themselves in them, but we know this is not the case, because P(Survival) = 1 has not been computed in a normal way like that. P(Survival) = 1 has been computed in a strange way that disconnects it from the specificity of the outcome.

You can only ever know which AI laboratory you are likely to be in if you have more information that is not provided in this description of the scenario. Maybe you know that the coin was biased, or the referee of the coin toss was corrupt, etc. However, nothing that happens in the course of the scenario, as described here, is going to tell you.

Normally "P(Outcome) = some very high value" would suggest an outcome with very low specificity: that there are many ways of obtaining that outcome, or many formally describable future situations in which that outcome has occurred. When we computed P(survival) = 1 this tricked us into thinking that the outcome had very low specificity, but this is really an illusion: the probability is not being made high due to any low specificity in the outcome. Rather, it is being made high because we think that an outcome with the same specificity as the one for the Type-2 laboratory will suffice to continue many more observers. Does this mean that we are wrong to say that P(Survival) = 1? Not necessarily, though some people would say that. The author suggests that this is a separate, more involved issue with a lot of philosophy needed. When it comes to the issue of knowing whether we are in a Type-1 AI laboratory or a Type-2 AI laboratory, we can ignore all this - and just realize that the outcome is equally specific in each case and that using P(Survival) = 1 to justify belief that you are in a Type-1 AI laboratory would be an error because it would be ignoring the fact that there is a disconnect between the probability and the specificity of the outcome – which is very unusual when applying the Bayesian method.

6.3 Analogy 3: An Example of How P(Survival) = 1 Can Become Disconnected From Any *Real* Low Specificity in the Reference Class.

Suppose there are two bags, each containing a mixture of 100 red and green balls.

Bag 1 contains 50 red balls and 50 green balls.
Bag 2 contains 25 red balls and 75 green balls.

A coin is tossed to select a bag: “heads” for Bag 1 and “tails” for Bag 2. You are not told the result of the coin toss. You are blindfolded and the bag that was selected is given to you. You take a ball out of the bag. The bag is then taken away. You are left holding the ball that you took from the bag. You remove the blindfold.

The ball is red. What is the probability that the coin toss result was “heads”?

Because of the greater number of red balls in Bag 1, it is obviously more likely that the coin toss result was “heads”, and that you were given Bag 1. You could perform the Bayesian calculation, as discussed previously, to obtain a probability for this.

The Bayesian calculation is entirely consistent with the idea of looking at the reference class of possible situations that are consistent with your experience after the experiment: in fact, the Bayesian calculation is just a shortcut to achieving the same result. The possible situations in the reference class could be represented with all kinds of details, such as the room temperature, the colour of the floor, etc., but we can ignore all that and just represent each situation in terms of which bag was given to you and which ball was selected from the bag. We can imagine that each ball in each bag has a number from 1 to 100.

When you have taken the ball from the bag, but have not yet removed the blindfold, your reference class of possible situations will consist of situations in which you were given Bag 1 and situations in which you were given Bag 2.

There will be 100 situations in which you were given Bag 1: one situation for each ball (1-100) that could have been taken from the bag. Bag 1 contains 50 red balls and 50 green balls, so there will be 50 situations in which you were given Bag 1 and took a red ball from the bag and 50 situations in which you were given Bag 1 and took a green ball.

Similarly, there will be 100 situations in which you were given Bag 2: one situation for each ball (1-100) that could have been taken from the bag. Bag 2 contains 25 red balls and 75 green balls, so there will be 25 situations in which you were given Bag 2 and took a red ball from the bag and 75 situations in which you were given Bag 2 and took a green ball.

You now remove the blindfold and look at the ball that you took from the bag. You see that it was a red ball. You can now update your reference class with this knowledge, removing all the situations in which you did not take a red ball from the bag. The 1 situation in which you were given Bag 1 and took a green ball is removed, and the 99 situations in which you were given Bag 2 and took a green ball are removed.

The reference class now contains the following situations:

50 situations in which you were given Bag 1 and took a red ball from the bag
25 situations in which you were given Bag 2 and took a red ball from the bag

There is a total of 75 situations, and 50 of these are ones in which you were given Bag 1. Therefore:

P(You were given Bag 1) = $^{50}/_{75}$ = $^{2}/_{3}$ = 0.667

i.e. observing that you took a red ball from the bag has caused the probability that you were given Bag 1 to increase from 0.5 to 0.667.

This is entirely consistent with the Bayesian calculation:

P_P(Bag 1) = 0.5
P_P(Bag 2) = 0.5
P(Bag 1 | Red Ball) = $^{50}/_{100}$ = 0.5
P(Bag 2 | Red Ball) = $^{25}/_{100}$ = 0.25

$$P(\text{Bag 1}) = \frac{P_P(\text{Bag 1}) \times P(\text{Bag 1} \mid \text{Red Ball})}{P_P(\text{Bag 1}) \times P(\text{Bag 1} \mid \text{Red Ball}) + P_P(\text{Bag 2}) \times P(\text{Bag 2} \mid \text{Red Ball})}$$

$$= \frac{0.5 \times 0.5}{(0.5 \times 0.5) + (0.5 \times 0.25)}$$

P(Bag 1) = $^{2}/_{3}$ = 0.667

which agrees with the result that was obtained with the reference class approach.

Suppose now that you want to make it more likely that a red ball will be drawn from Bag 1: you want to increase P(Bag 1 | Red Ball). Bag 1 currently contains 50 red balls and 50 green balls, but you replace 25 of the green balls with red balls, so that Bag 1 now contains 75 red balls and 25 green balls.

There will be 100 situations in which you were given Bag 1: one situation for each ball (1-100) that could have been taken from the bag. Bag 1 contains 75 red balls and 25 green balls, so there will be 75 situations in which you were given Bag 1 and took a red ball from the bag and 25 situations in which you were given Bag 1 and took a green ball.

The situation for Bag 2 is the same as before. There will be 100 situations in which you were given Bag 2: one situation for each ball (1-100) that could have been taken from the bag. Bag 2 contains 25 red balls and 75 green balls, so there will be 25 situations in which you were given Bag 2 and took a red ball from the bag and 75 situations in which you were given Bag 2 and took a green ball.

As before, you take a ball from the bag and find that it is a red ball. You remove the situations from the reference class in which you did not take a red ball from the bag.

The reference class now contains the following situations:

75 situations in which you were given Bag 1 and took a red ball from the bag
25 situations in which you were given Bag 2 and took a red ball from the bag

There is a total of 100 situations, and 75 of these are ones in which you were given Bag 1. Therefore:

P(You were given Bag 1) = $^{75}/_{100}$ = $^{2}/_{3}$ = 0.75

i.e. observing that you took a red ball from the bag has caused the probability that you were given Bag 1 to increase from 0.5 to 0.75.

When there were 50 red balls and 50 green balls in Bag 1, taking a red ball from the bag caused the probability that you were given Bag 1 to increase from 0.5 to 0.667, but when there are 75 red balls and 25 green balls in Bag 1, taking a red ball from Bag 1 caused the probability that you were given Bag 1 to increase from 0.5 to 0.75: changing some of the red balls in Bag 1 into green balls means that taking a red ball from Bag 1 causes a greater increase in the probability that you were given Bag 1.

This is consistent with the Bayesian calculation:

P_P(Bag 1) = 0.5
P_P(Bag 2) = 0.5
P(Bag 1 | Red Ball) = $^{75}/_{100}$ = 0.75
P(Bag 2 | Red Ball) = $^{25}/_{100}$ = 0.25

$$P(\text{Bag 1}) = \frac{P_P(\text{Bag 1}) \times P(\text{Bag 1} \mid \text{Red Ball})}{P_P(\text{Bag 1}) \times P(\text{Bag 1} \mid \text{Red Ball}) + P_P(\text{Bag 2}) \times P(\text{Bag 2} \mid \text{Red Ball})}$$

$$= \frac{0.5 \times 0.75}{(0.5 \times 0.75) + (0.5 \times 0.75)}$$

P(Bag 1) = ¾ = 0.75

Why did changing green balls into red balls in Bag 1 have this effect? It increased the value of P(Bag 1 | Red Ball), but this should really be viewed in reference class terms. Changing green balls into red balls gave more ways of taking a red ball from Bag 1, so it added situations to the reference class in which Bag 1 had been given to you and you had taken a red ball from the bag. These situations were then combined with the other situations in which a red ball had been taken from the bag to give the final reference class and the final probability – and all this is reflected in the Bayesian calculation.

Suppose now that MWI is actually true, and also suppose that the selection of a ball from the bag can be treated as if it is a quantum event: you can treat it as if the world

splits and there is a separate world for each ball. It may seem strange to do this, as this is clearly all at a high enough level that it should be viewed classically, but this is an analogy rather than an attempt to represent the world accurately. You start again with the original situation in which Bag 1 has 50 red balls and 50 green balls and Bag 2 has 25 red balls and 75 green balls.

You try a more unusual method of increasing P(Bag 1 | Red Ball): You set up a mechanism which kills you immediately if you are given Bag 1 and then take a green ball from the bag. There is no such mechanism for Bag 2: if you are given Bag 2 you are safe, no matter what colour of ball you take out of it.

You reason that if you are given Bag 1, and take a green ball out of it, you will be killed immediately, so you will never be able to observe that you have taken a green ball out of Bag 1. Therefore, you can be sure of taking a red ball out of Bag 1:

P(Bag 1 | Red Ball) = 1

People can object to this by saying that this is not like a proper quantum suicide experiment, because the situation here would never allow it to be quick enough, but this would be missing the point. We can assume that the addition of a mechanism to kill yourself to the experimental procedure causes P(Bag 1 | Red Ball) = 1 *for the sake of argument*. This is not about whether or not P(Bag 1 | Red Ball) = 1 is valid: it is about how setting yourself up to be killed does not tell you anything *even if it is valid to say that P(Bag 1 | Red Ball) = 1*.

Suppose you are given a bag – you do not know which one – and you take a ball out of it. It turns out to be a red ball.

You do the Bayesian calculation:

P_P(Bag 1) = 0.5
P_P(Bag 2) = 0.5
P(Bag 1 | Red Ball) = 1
P(Bag 2 | Red Ball) = $^{25}/_{100}$ = 0.25

$$P(\text{Bag 1}) = \frac{P_P(\text{Bag 1}) \times P(\text{Bag 1} \mid \text{Red Ball})}{P_P(\text{Bag 1}) \times P(\text{Bag 1} \mid \text{Red Ball}) + P_P(\text{Bag 2}) \times P(\text{Bag 2} \mid \text{Red Ball})}$$

$$= \frac{0.5 \times 1}{(0.5 \times 1) + (0.5 \times 0.25)}$$

P(Bag 1) = $^{4}/_{5}$ = 0.8

All else being equal, assuming MWI and ensuring that you are terminated immediately if you take a red ball from the bag seems to increase the probability that you have been given Bag 1 from 0.667 to 0.8.

But is this really the case? The Bayesian formula can be explained by the reference class approach that we have been using. What does a reference class consideration tell you?

When you are holding a ball, and have not yet looked to see what colour it is, the reference class contains 50 possible situations in which you have been given Bag 1 and taken a red ball out of it, 25 situations in which you have been given Bag 2 and taken a red ball out of it and 75 situations in which you have been given Bag 1 and have taken a green ball out of it. The 50 situations in which you have been given Bag 1 and have taken a green ball out of it are no longer in the reference class: you would have already been killed if this had happened.

It should be obvious here that setting up the mechanism to kill you has not altered the reference class in any way that will matter if you find that you have taken a red ball from the bag. The 50 situations in which you have been given Bag 1 and have taken a green ball out of it have been removed, due to the mechanism to kill you, but these would have been removed from the reference class when you found out you had a red ball anyway.

When you find out that you have taken a red ball from the bag, your reference class becomes 50 situations in which you have been given Bag 1 and have taken a red ball out of it and 25 situations in which you have been given Bag 2 and have taken a red ball out of it. There are 75 situations, and you have been given Bag 1 in 50 of them, so the proportion of possible situations in which you were given Bag 1 is 50/75 = 2/3 = 0.667. So:

P(You were given Bag 1) = $^{50}/_{75} = {}^{2}/_{3} = 0.667$

This is the same result that was obtained *without* the mechanism to kill you in place. In fact, what has been done with the reference class above is almost identical: the only difference is that with the mechanism to kill you in place, the situations in which you were given Bag 1 and have taken a green ball are never allowed into the reference class – rather then being allowed in and removed when you observed that you have taken a red ball.

Why has the Bayesian method failed here? The method used to make it certain that, if you had Bag 1 you would take a red ball – to ensure that P(Bag 1 | Red Ball) = 1, is a non-standard method of changing the probability. Usually, we would put more red balls in Bag 1, or take some green balls out of Bag 1, or change some of the green balls in Bag 1 into red balls. All of these would actually affect the amount of representation in the reference class of the situations in which you were given Bag 1 and had taken a red ball out of it relative to the amount of representation of situations in which you were given

Bag 2 and had taken a red ball out of it. Alternatively, we might have taken some red balls out of Bag 2, or put more green balls into Bag 2, or changed some of the green balls in Bag 2 into red balls. This would reduce the representation of situations in the reference class corresponding to situations in which you were given Bag 2 and have taken a red ball. (We need to be careful here: if we add a lot of green balls to Bag 2, so that it contains more than 100 balls, while leaving the number of red balls the same, it may seem that we have not reduced the amount of representation of situations in which you were given Bag 2 and have taken a red ball out of it, but rather have simply added situations in which you were given Bag 2 and have taken a green ball. This would not be the case: the total degree of representation of the Bag 1 and Bag 2 situations would have to be the same if all you have done is add or remove balls from the bags, and therefore, as the numbers of balls in each bag are now different, you would have to consider things in terms of proportions of the reference class rather than just counting situations in the simple way that we have been doing here.) The method used to ensure that P(Bag 1 | Red Ball) = 1 does not rely on adding or removing red or green balls. Instead it works by removing a proportion of the situations from the reference class – by ensuring that you are not there to observe them – and these are situations that would be removed anyway on finding out that you had taken a red ball from the bag. This does nothing to increase the proportion of situations that remain, in which you were given Bag 1 and have taken a red ball out of it, relative to the proportion of situations in which you were given Bag 2 and have taken a red ball out of it – which is all that matters.

One way of viewing this is in terms of the *history of situations* being non-standard when you try to use your possible death to manipulate the reference class. The idea is to reduce the specificity of situations in which you take a red ball out of Bag 1. If you do this by altering the numbers of balls in bags then you change what has to happen for a ball of some colour to be taken from a bag: you can require it to have a more specific past. For example, if Bag 1 contains 50 red balls and 50 green balls, then the actual physical process of a red ball being taken from the bag has to be one of someone's hand going into the bag and selecting one of the 50 red balls among the 50 green balls. If another 25 red balls are added to the bag then the hand going into the bag has to select one of 75 red balls with the 25 green balls also in the bag. The history of a situation in which a red ball has been taken from the bag has been altered to make it less specific, meaning that there can be a greater representation of such situations in the reference class. When you set up a mechanism to kill you if a green ball is taken out of the bag, on the other hand, this cannot do anything to alter the history of a situation in which a red ball is taken out of the bag: the mechanism does not even activate unless a green ball is taken out of the bag, and so cannot feature in the history of a situation in which a red ball is taken from the bag in any important way. (It can feature in the situation just by being there, inactive, waiting to be triggered, but in that inactive state it would have as much influence on the reference class as a houseplant.)

7 Measure Reduction When Worlds Split in a Level IV Multiverse

7.1 How Measure would Decrease in a Level IV Multiverse

In this article, the position has been taken that when worlds split in MWI, if MWI is true, there must be a reduction of measure for situations in which an observer can exist – and it has been shown that this must be the case. This means that, if MWI is true, splitting of worlds in MWI, before you know the outcome, must be associated with a reduction in the measure of *observer moments*.

MWI is a theory describing a Level III multiverse, according to the classification system devised by Max Tegmark (Tegmark, 2003). Max goes even further than this, however, proposing a much more expansive multiverse, called the *mathematical universe* consisting of every describable mathematical object (Tegmark, 1998, 2003, 2007). Max classifies such a multiverse as a Level IV multiverse.

The Level III multiverse theory of MWI and the Level IV mathematical universe theory do not contradict each other. If reality is a Level IV multiverse then it would contain lots of different "worlds" inhabited by different observers. We could be living in a Level III MWI multiverse, which is just one of the "places" in the more expansive Level IV multiverse.

This article has argued that a measure reduction must occur in the reference class when worlds split in MWI, before we know of the outcome, and if MWI is true and we are also in a Level IV multiverse, the measure for observer moments in the Level IV multiverse should match up with the way that measure of observer moments is reduced as worlds split.

An explanation of this is not too difficult. In a Level IV multiverse, all objects will not be equal: objects will exist with varying measure. An indication of the measure of an object is its specificity – and we can view this as being the amount of information it contains. In the author's own writing on very expansive multiverses, it has been argued that the measure of an object will correspond to the proportion of algorithms that will produce that object by interpretation of other objects (Almond, 2010). It may be that an even more general position than this has to be taken, and that the author has merely been discussing a special case: something like Max's mathematical universe is possibly relevant here. In any event, in any plausible version of this kind of theory, the greater the amount of information needed to describe an object, the less common that object should be in the multiverse, because whatever the way is in which such an object is implied to exist, the greater information content will mean that the ways in this specific object can exist will be a smaller proportion of all the ways in which objects in general can exist. In general, the more specific an object is – the greater its information content

is – the lower its measure should be. Any proposal for anything like a Level IV multiverse or modal realism is going to have to say that, all else being equal, the information content of an object is related to its measure: otherwise you would have a multiverse in which dinosaurs stamping on branches had the same measure as hydrogen nuclei fusing together, when dinosaur-branch-stampings are clearly demanding a lot more out of reality than fusion events, due to their higher information content – their higher specificity. All else being equal, specificity must imply less measure and, in fact, it would be hard to construct an ontology in which it did not.

Whenever a world splits in MWI, each new world is one in which an outcome of a quantum event has a different world. Each world which has split off from the previous world therefore contains the information about how the event happened in that particular world. That is to say, if there is a quantum event with possible outcomes of 0 and 1, two worlds will be produced. In one world the event has happened with an outcome of 0, and this means that this world now has an extra bit in its description, indicating that the outcome was 0, that was not in the world before the split. Likewise, in the other world the event has happened with an outcome of 1, and this needs to be added to the world's description, adding an extra bit to it. When a world splits into two worlds, therefore, each world gains one bit of information. As splitting continues over time, worlds gather progressively more information.

In a Level IV multiverse, there is no "flow" of time in the fundamental sense of the idea as most people think of it. Objects corresponding to these worlds at different times could be regarded as existing in an atemporal way, so the multiverse would contain an object corresponding to the state of the world just before the splitting, and it would contain objects corresponding to the states of the world with the different outcomes of the quantum event just after the splitting. In the case of a quantum event with two outcomes, 0 and 1, we can think in terms of an object corresponding to the state of the world just before the split and two objects, each corresponding to the state of the world just after the split in which the outcome is different: one such object would correspond to a world in which the outcome had been 0 and in the other would correspond to an object in which the outcome had been 1.

The descriptions of the two objects corresponding to the state of the world just after the split need to include information about what the outcome of the quantum event was in a world in that state: each description needs to say whether the outcome was 0 or 1 in that world. This requires one bit of information in the description, which is not required in the description of the state of the world before the split. Therefore, splitting gives rise to states with greater information content in their descriptions. As has already been discussed, the greater the information content in an object, the lower its measure should tend to be, so each of the two objects corresponding to the two states for the world just after the splitting should have lower measure than the object corresponding to the state of the world just before the split: as worlds split, the measure of objects corresponding to the state of the world in the Level IV multiverse should be reduced.

Total measure, however, would be conserved in "normal" situations: when a world splits into two worlds, each of the two "after" world states will have a measure which is half that of the "before" world state, with the measure of both "after" world states combined being the same as the measure just before the split.

There remains the issue of actually relating this to observer moments, rather than just parts of the wave function. In the author's view, your mental experiences as an observer are due to an interpretation or abstraction of your underlying brain, which is itself an interpretation or abstraction of the underlying brain cells in it, which is an interpretation or abstraction of underlying components of cells, etc. down through molecules, atoms to the quantum wave function. If you are in a Level III, MWI multiverse within a Level IV multiverse, it should make sense to think that the splitting of worlds corresponds to the physical substrate underlying your mental existence undergoing a decrease in measure, in the way that has been discussed, which will lead to a corresponding decrease in the measure of anything, including your mental states, that is implied by, or abstracted from, that substrate.

7.2 Distinguishing Between MWI and non-MWI in a Level IV Multiverse

If we do live in a Level IV multiverse, then this is not in itself equivalent to MWI being true: MWI is a specific quantum mechanics theory about the wave function splitting. We might imagine living in some "world" in a Level IV multiverse in which we should regard MWI as true for us, but we might also imagine living in a Level IV multiverse in which we should regard MWI as not true for us: this would not mean that MWI was not true "somewhere else" in the multiverse, but it would just mean that it was not true in our part of it. We can consider the idea of quantum suicide in such a context, to see if changes anything about the idea that surviving quantum suicide would suggest that MWI is true. In this context, the question would be one of whether surviving quantum suicide suggests that you are any more likely to be in a region of the Level IV multiverse where MWI is true.

The author should be clear about the reasons for discussing this. Not everybody takes the idea that reality is a Level IV multiverse seriously. Readers familiar with the author's previous writing on the subject will know that the author *does* take this kind of proposal seriously. It is not the author's intention, however, to try to persuade readers of that here. Rather, putting the situation into the context of a Level IV multiverse, where the possible worlds that we are considering become actual worlds, will be a good way of showing how the decrease in measure when worlds split in MWI prevents us from using quantum suicide to find out whether MWI is true. Essentially, the argument that we have been using can be mapped onto a Level IV multiverse, and by seeing why surviving quantum suicide tells you nothing in such a context, it should be more clear why the same problem arises when we map the situation from a Level IV multiverse of actual worlds back onto the probability distribution of possible worlds.

Suppose we accept that a Level IV multiverse exists. Any world which can be formally described actually exists. All issues of science are now, then, issues of locating ourselves in such a multiverse.

Suppose there are some parts of this Level IV multiverse corresponding to situations in which MWI is locally true. Max classifies a multiverse in which MWI is true as a Level III multiverse, so what we are supposing here is that some parts of the Level IV multiverse correspond to Level III multiverses. We will call these parts of the Level IV multiverse *Type-1 worlds*.

Suppose also that there are some parts of the Level IV multiverse corresponding just to "Copenhagen" type worlds. These worlds actually have laws of physics in which wave function collapse actually occurs and, as far as anyone in them is concerned, MWI is not true. (Of course, in the context of a Level IV multiverse, this means that MWI is not *locally* true.) We will call these *Type-2 worlds*.

There will be an infinity of Type-1 worlds and an infinity of Type-2 type worlds, but each will exist with some measure.

Suppose you know that you are living in either a Type-1 world or a Type-2 world: in other words, either MWI is true for you or it is not. Suppose you are given a machine called an "ultimate analyzer" which can answer any questions about the relative measures of objects in the Level IV multiverse.

You ask the ultimate analyzer what the relative measure is of observer moments which, to you, would be indistinguishable from the one you are having right now, in both Type-1 worlds and Type-2 worlds. It tells you that the measure is the same in each case. This is of no help at all: it means you are equally likely to be in a Type-2 world or a Type-1 world. Now, a good argument can be made, using the same kind of reasoning that was used earlier, in 3.4, that this is not a likely scenario: that the ultimate analyzer would probably report that the measure of observer moments like yours in Type-1 worlds is actually higher, because the situation of such observers is one that needs no wave function collapse mechanism in its physics, and so is less specific and more "common" in the Level IV multiverse. For simplicity, however, we will assume for now that you have no way of distinguishing between types of world.

You plan a sequence of quantum suicide experiments. You think that if you are in a Type-1 world, P(Survival) = 1, whereas it will be much less in a Type-2 world. You therefore think that, if you survive, this will be evidence that you are in a Type-1 world.

Later, you find yourself alive with a memory of having done the above probability calculations and performing the experiments. You know that you have survived the quantum suicide experiments. You are about to conclude that you are probably in a Type-1 world, and that MWI is probably true for you, but you pause. You ask the ultimate analyzer to compute the relative measures of observer moments

indistinguishable from the one you are having now in both Type-1 worlds and Type-2 worlds. *It tells you that the measure is still the same: your experiment did not change the relative measure of the observer moments*. This is because, in both cases, the post-experiment measure of observer moments is reduced equally for each type of world, so the relative measure of observer moments in each stays the same. An observer moment of a version of you who survived in a Type-1 world has the same measure as that of a version of you who survived in a Type-2 world. Your experiment told you nothing about which theory is more likely to be true.

Earlier, it was supposed that, before performing the quantum suicide experiments, the ultimate analyzer told you that the relative measure is the same for each kind of observer moment, but suppose this is not the case? The ultimate analyzer might tell you, before performing any experiments, that observer moments in World-1 type worlds have greater measure, or that observer moments in World-2 type worlds have greater measure. It should be apparent that, in this situation, performing quantum suicide experiments is not going to change this. The quantum suicide experiments will cause the same proportionate reduction in measure for observer moments in each type of world, and the situation, with regard to relative measure of observer moments in each type of world will be just as it was before the experiments were performed. Your view of the type of world in which you are likely to exist, therefore, should be completely dependent on what the ultimate analyzer told you before you started the experiments: its answer will not change when you perform them.

Someone might complicate this by proposing that, even if you are in a Type-2 world, you might reasonably think that the survival of a version of you in a Type-1 world will provide you with continuity and that you should think that P(Survival) = 1. We would be into deep issues here. That would not really help, though, because it still leaves no way of finding out where you are.

In any event, you may not believe that you are in a Level IV multiverse, and many people reading this will not, so what then?

Suppose there were some version of the ultimate analyzer that did not need a Level IV multiverse in which to work: instead, it could just tell you the relative measure of different kinds of observer moment in a probability distribution, and it would work by doing computations about possible worlds rather than actual worlds. The author suggests that this kind of ultimate analyzer and the ultimate analyzer needed to work in a Level IV multiverse should actually work the same way: they would be the same device. In a Level IV multiverse, modal realism would be true, and this would merely mean that everything in a probability distribution that you could compute would actually be real. Even if you are not in a Level IV multiverse, you should still compute the measure of anything in a probability distribution as if you are in one, because the measures of things in the probability distribution will be the same as the ones for a Level IV multiverse. Very specific things which can only occur in a small number of ways will

tend to have low measure and less specific things which can occur in many ways will tend to have higher measure. You can do statistics without worrying about whether modal realism, is true or not: if modal realism is true, all your probability calculations are about a “probability space" that is actually "real", and if modal realism is not true, all your probability calculations are about a probability space that does not actually exist – which is just being used as a conceptual device, but the structure of the space is the same. The author discussed this issue in a previous article, where what was called the *principle of modal realism equivalence* was introduced (Almond, 2008).

All this means that, regardless of whether you are in a Level IV multiverse or not, if you ask the ultimate analyzer to compute a measure for Type-1 world and Type-2 world observer moments, and then perform some quantum suicide experiments, and survive them, before asking the ultimate analyzer again, the measure will have decreased in both cases, so that the relative measure remains the same: you will know nothing.

Incidentally, a good argument can be made that modal realism should be preferred because probability, if properly considered, actually requires us to make us of a conceptual structure which is exactly the same as a Level IV multiverse – and that should be a really big giveaway. As an analogy, if you find yourself wandering around a dark building and you have to make use of some conceptual model of a library, it would normally be a good indication that you are in one. One day, people might think it was obvious that the conceptual probability space we use now in calculations was nothing more than a description of a Level IV multiverse around us, and we were effectively using a map of it and all but assuming its existence whenever we did probability calculations.

Ultimately, all this puts us back with the "ultimate reference class" argument given previously: the ultimate reference class argument is just a version of the argument just given, without worrying about whether we are already in a Level IV multiverse. Quantum suicide will not tell us anything that we did not already know. It will not even change probabilities, because we will start with some probability distribution of different situations in which we could be, which will feature both MWI and non-MWI explanations, and the quantum suicide experiments will affect both kinds of probability equally. In a Level IV multiverse we should be able to see what happens easily, but even without a Level IV multiverse our probability distribution is going to be constructed and handled in the same way.

8 Conclusion

Max Tegmark has argued that, by performing repeated quantum suicide experiments, an observer can gain evidence that the many-worlds interpretation of quantum mechanics (MWI) is true. There is a Bayesian justification for this. The idea is that if you arrange for yourself to be terminated if a particular outcome occurs, and for yourself to survive if another outcome occurs, if MWI is true, both outcomes can be regarded as occurring, yet you can expect to be certain of observing the outcome corresponding to the future in which you survive, because you cannot observe the outcome corresponding to the future in which you are dead. In a long sequence of experiments, you could be assured of survival if MWI is true, but if MWI is not true, the chance of death in each experiment would make your chance of surviving until the end small. If you find yourself alive at the end of the sequence of experiments, you know that this was certain if MWI is true and very unlikely if MWI is not true. You can therefore perform a Bayesian calculation which will adjust the prior probability you assigned to MWI being true, increasing the probability that MWI is true significantly. In fact, no matter how low your prior probability of MWI being true is, surviving a long enough sequence of quantum suicide experiments should allow you to conclude that MWI is true with any degree of confidence that you like.

This idea has problems. In a previous article, *Quantum suicide would not suggest that MWI is correct – even to the person doing it*, available at http://www.paul-almond.com/QuantumSuicide.pdf or http://www.paul-almond.com/QuantumSuicide.doc, the author argued that this view is incorrect and that repeated quantum suicide experiments will indicate nothing about the likelihood of the many-worlds interpretation being correct, even to the observer who is undergoing them. The argument was that measure decreases as worlds split in MWI, and that this reduction in measure makes it just as implausible, if MWI is true, that you are in an observer moment after a quantum suicide experiment, rather than in one before it, as it would be to think that you survived a quantum suicide experiment in a reality in which MWI is not true.

Max asked two challenging questions about this article. The first question involved asking for justification for the choice of reference class. Why have a reference class consisting of before and after observer moments? This is a potentially awkward issue. The author's view is that the choice of reference class is justified, and some justification for it has been given in this article. Nevertheless, it allows for possible controversy which should be avoided. The second question involved asking what the specific flaw was in the Bayesian calculation used to justify the idea that surviving quantum suicide would suggest that MWI is correct.

In this article, an improved version of the argument has been presented. The improved argument uses a reference class which should be less controversial. A reference class of

descriptions of possible candidates for your situation at any time is used, which is not as subject to challenges about reference class as the reference class in the previous argument. The reference class is considered represented in two ways.

In the first way of considering the reference class, it is considered as the *ultimate reference class*: the set of all formal descriptions of possible situations that are consistent with what you know about your situation right now, with some maximum description length of n bits. This can also be described as the set of all formal descriptions of *you-centred possible worlds* with description lengths of n bits or less. A you-centred possible world is a possible world that is consistent with what you know about your situation right now. If you consider the ultimate reference class of possible situations in which you have survived any number of quantum suicide experiments, the proportion of the ultimate reference class corresponding to situations in which MWI is true will be no different than it was before the quantum suicide experiments were performed.

In the second way of considering the reference class, it is viewed in terms of "specificity" – lack of specificity corresponding with the number of ways in which situations can exist, in the way in which simple probability calculations are generally performed. The result is the same: surviving quantum suicide does not change the proportion of situations in the reference class in which MWI is true. Situations in which MWI is true and you survive and MWI is not true and you survive are equally specific.

This can also be viewed in terms of reduction of measure. When a quantum event occurs, and you do not know the outcome, each outcome gets a share of the measure: measure is conserved and divided up among the outcomes. This will generally be seen as occurring when MWI is not true, but it also occurs when MWI is true, so that the same reduction in measure occurs for both MWI being true and MWI not being true. When the outcome becomes known, only MWI and non-MWI situations for that particular outcome remain in the new reference class, and these will have been through an identical process of measure reduction, so that when they occupy the entire reference class, nothing has changed.

Without an actual explanation of what is wrong with the Bayesian calculation, it may seem that there are two opposing arguments – the argument using a reference class consideration to show that surviving quantum suicide would *not* suggest that MWI is true and the argument relying on the Bayesian calculation to show that surviving quantum suicide *would* suggest that MWI is true – and that both arguments might have merit, so that it may not be clear which argument should be regarded as correct. It would be incorrect to think this. It has been shown that the Bayesian calculation is merely a shortcut for the reference class considerations discussed here and, in most situations that we could imagine, the Bayesian calculation would agree with the reference class consideration. With quantum suicide, however, the situation is not like

most situations in which the Bayesian calculation is used. Usually, increasing the probability of an outcome is done in a way which actually increases its representation in the reference class, relative to the representation of other outcomes, but this is not what is being done with quantum suicide: instead, the unwanted outcomes are being removed from the reference class. From the point of view of an observer about to undergo a quantum suicide experiment, this may have the appearance of an increase in probability of survival – and this may be a correct view, although it will be controversial – but there is no increase in measure for the situation in which MWI is true and you survive relative to that of other situations – and in particular relative to that of the kind of situation in which MWI is *not* true and the same outcome occurs: the probability has become disconnected from the specificity, which means that the probability value is not a true indication of specificity in the reference class and the Bayesian method cannot be relied upon anymore. A further issue is that the increase in the probability of survival is achieved without any special effect on the history of the kind of situation that it is supposed to be making more probable, which again is non-standard. Another issue is that the way in which the probability of 1 is assigned to the survival outcome when MWI is true is a non-standard way of assigning probability. When you give one outcome a probability of 1 because you are not going to be in a state that allows you to observe another outcome, the outcome getting the probability of 1 is not getting it because of its own lack of specificity: it is getting it because of the state that *you* are in when another outcome occurs. It is all about you – not the outcome that is getting all the probability. This is important because the specificity of that outcome is not being measured against the outcome that you are not observing: it is being measured against the corresponding outcome in which the hypothesis, in this case MWI, is not true. These issues should not be regarded as an argument against assigning a probability of 1 to survival in suicide experiments when MWI is true: rather, they should be considered as an explanation of how the non-standard features of quantum suicide cause the Bayesian calculation departs from the more reliable reference class consideration. The idea that the Bayesian calculation justifies the view that surviving quantum suicide suggests that MWI is incorrect.

Surviving quantum suicide would tell you nothing about the likelihood of MWI being true.

9 Acknowledgements

Max Tegmark has been helpful in discussions relating to this article. Max's help was in asking challenging questions, after the author had made the previous version of this argument, which persuaded the author that the previous argument was not persuasive enough in two respects and that an improved version was needed which was stronger in these two respects. These questions were asked by Max purely to challenge the argument, and their influence has been positive in helping with the formulation of the stronger version of the argument given here. Yvonne Deborah Finch and Darla Lundell has also been helpful in discussions.

10 References

Almond, P., 2008. *The Principle of Modal Realism Equivalence*. [Online] paul-almond.com. Available at: http://www.paul-almond.com/ModalRealismEquivalence.pdf, http://www.paul-almond.com/ModalRealismEquivalence.doc or http://www.paul-almond.com/ModalRealismEquivalence.htm [Accessed 11 February 2011].

Almond, P., 2010. *Minds, Substrate, Measure and Value, Part 4: The Cosmological Many-Interpretations View*. [Online] paul-almond.com. Available at: http://www.paulalmond.com/Substrate4.pdf or http://www.paul-almond.com/Substrate4.doc [Accessed 19 February 2011].

Almond, P., 2010. *Repeated quantum suicide would not suggest that MWI is correct – even to the person doing it*. [Online] paul-almond.com. Available at: http://www.paul-almond.com/QuantumSuicide.pdf or http://www.paul-almond.com/QuantumSuicide.doc [Accessed 11 February 2011]. (This is the previous version of the argument of which a stronger version is given in this article.)

Almond, P., 2011. *The self-indication assumption almost stops the Doomsday argument. Almost. V2.0*. [Online] paul-almond.com. Available at: http://www.paul-almond.com/QuantumSIA2.pdf or http://www.paul-almond.com/SIA2.doc [Accessed 20 February 2011].

Bostrom, N., 2005. *Self-Location and Observation Selection Theory*. [Online] The Anthropic Principle. Available at: http://anthropic-principle.com/preprints/self-location.html [Accessed 8 December 2010].

Carter, B., 1983. The anthropic principle and its implications for biological evolution. *Philosophical Transactions of the Royal Society of London*, A310, pp.347-363.

Elga, A., 2000. Self-locating belief and the Sleeping Beauty problem. *Analysis*, 60(2), pp.143-147. (Also available at: http://www.princeton.edu/~adame/papers/sleeping/sleeping.html [Accessed 10 December 2010]). (The problem was discussed earlier in unpublished work by Arnold Zuloff.)

Everett, H., 1957. Relative State Formulation of Quantum Mechanics. *Reviews of Modern Physics,* 29, pp.454-462.

Gott, J. G. III, 1993. Implications of the Copernican principle for our future prospects. *Nature*, 363, pp.315-319.

Marchal, B., 1988. Mechanism and Personal Identity. *Proceedings of WOCFAI 91* (Paris. Angkor), pp.335-345. (Also available at: http://iridia.ulb.ac.be/~marchal/publications/M&PI_15-MAI-91.pdf [Accessed 10 December 2010].)

Moravec, H., 1988. The Doomsday Device. *Mind Children: The Future of Robot and Human Intelligence*. Harvard: Harvard University Press. p.188. (Also available at: http://books.google.com/books?id=56mb7XuSx3QC&lpg=PP188&pg=PA188 [Accessed 10 December 2010].)

Price, M. C., 1995. *The Many-Worlds FAQ*. [Online] The Anthropic Principle. Available at: http://www.anthropic-principle.com/preprints/manyworlds.html [Accessed 12 December 2010]. (Also available at: http://www.hedweb.com/everett/everett.htm [Accessed 12 December 2010] and http://kuoi.com/~kamikaze/doc/many-worlds-faq.html [Accessed 12 December 2010].)

Tegmark, M., 1997. *The Interpretation of Quantum Mechanics: Many Worlds or Many Words?*. [Online] arXiv:quant-ph/9709032. Available at: http://arxiv.org/abs/quant-ph/9709032 [Accessed 8 December 2010]. (Also available at: http://xxx.lanl.gov/PS_cache/quant-ph/pdf/9709/9709032v1.pdf [Accessed 12 February 2011].)

Tegmark, M., 1998. Is the theory of everything merely the ultimate ensemble theory? *Annals of Physics*, 270, pp.1-51. (Also available online at: http://arxiv.org/PS_cache/gr-qc/pdf/9704/9704009v2.pdf [Accessed 17 February 2011]).

Tegmark, M., 2003. Parallel Universes. *Scientific American*, May 2003, pp.40-51. (Also available online at: http://space.mit.edu/home/tegmark/PDF/multiverse_sciam.pdf [Accessed 17 February 2011].)

Tegmark, M., 2007. The Mathematical Universe. *Found.Phys*, 38, pp.101-150. (Also available online at: http://arxiv.org/abs/0704.0646 [Accessed 5 December 2010]).

Yudkowsky, E., 2008. *The Dilemma: Science or Bayes*. [Online] paul-almond.com. Available at: http://lesswrong.com/lw/qa/the_dilemma_science_or_bayes/ [Accessed 22 February 2011].