Candidate name: Jura Ivankovic

Title: Self-locating credences

Word count: 25,242

Abstract

In this essay, I will argue for a theory of self-locating credences based on Nick Bostrom's Observation Equation. A theory of self-locating credences should constrain a rational observer's credence function given their prior credences in non-self-locating propositions and their beliefs about how many observers exist in the world and what kind of evidence they have. Two notable theories of this kind have been proposed: one by Frank Arntzenius and Cian Dorr, and another by Nick Bostrom. In the essay, I argue that Arntzenius and Dorr's theory of self-locating credences has strongly counter-intuitive results when applied to certain kinds of cases, as Nick Bostrom has shown. On the other hand, Bostrom's theory is incomplete—Bostrom says that observers should be separated into reference classes in order to avoid the results that a theory like Arntzenius and Dorr's gives, but does not say anything about how observers should be separated into reference classes. In this essay, I develop a principled reference class definition to complement Bostrom's theory. I first give a preliminary reference class definition that is adequate in a special case—namely, whenever every observer knows what reference class they are in according to this preliminary definition. I then formulate a general reference class definition that I claim works generally, even when observers do not know what reference class they are in according to the preliminary definition. Following this, I apply the reference class definition to the Sleeping Beauty problem and conclude that her credence in heads should be 1/2 throughout. In the last part of the essay, I defend the theory from some potential objections.

Table of contents

Introduction………..4

Arntzenius and Dorr's theory……………………………………………………………………………………………6

Problems for Arntzenius and Dorr……………………………………………………………………………………..17

Bostrom's theory………22

Problems for Bostrom's theory…………………………………………………………………………………………30

My theory………37

Preliminary reference class definition……………………………………………………………………………….39

General reference class definition…………………………………………………………………………………….51

Explanation of the general reference class definition…………………………………………………………..56

Sleeping Beauty………61

Manley's objections………………………………………………………………………………………………………..66

Conclusion…….81

Bibliography……….83

Introduction

This essay is about self-locating credences. Let us begin with the following thought-experiment from 'Self-locating Priors and Cosmological Measures' by Frank Arntzenius and Cian Dorr:

> **Incubator:** The incubator is a construction consisting of a hundred cells, all identical on the inside and each with a door that is numbered #1–#100 on the outside. A fair coin is tossed. If it lands heads, 100 observers are created, one in each cell. If it lands tails, one observer is created in cell #1. Every possible observer knows this setup, and there are no other observers in the universe. An observer wakes up in their cell. Later, the observer exits their cell and observes that they were in cell #1.[1]

What should an observer's credence in heads be? There are several answers one might naturally give. One might think that, since the objective chance of the outcome of a coin toss being heads, the observer's credence in heads should be 1/2 throughout. One might think that the observer's credence in heads should be initially high because there are many more observers in the heads world, but that it should drop to 1/2 when the observer finds out that they were in cell #1. Or, one might think that the observer's credence in heads should initially be 1/2, but that it should rise to a much higher number when the observer finds out that they were in cell #1.

A theory of self-locating credences should tell us how an observer's prior credences in non-self-locating propositions, including especially beliefs about the number of observers in the world and the kinds of evidence they have, should constrain the observer's credences in self-locating propositions.

Two prominent such theories have been proposed: one by to Arntzenius and Dorr, and one by Nick Bostrom. In this essay, I will argue that Arntzenius and Dorr's theory must be

---

[1] Dorr and Arntzenius, 'Self-locating Priors and Cosmological Measures', p. 15.

incorrect given the strongly counterintuitive results that it implies about certain kinds of thought-experiments. On the other hand, I will argue that Bostrom's theory is incomplete. Bostrom's theory says that observer-moments should be divided into reference classes, but does not specify any rules about how this should be done. An observer's credences depend on how reference classes are defined, so the fact that Bostrom does not state how reference classes should be delineated means that his theory cannot give us answers about what an observer's credences should be in various controversial cases. The main thing I will do in this essay will be to formulate and argue for a reference class definition that can be used to make Bostrom's theory complete.

The broad plan of this essay will be as follows. First, I will describe Arntzenius and Dorr's theory of self-locating credences. Then, I will argue that Arntzenius and Dorr's theory cannot be correct. My main argument for this claim will be that Arntzenius and Dorr's theory makes strongly counterintuitive predictions in a wide range of thought-experiments. Second, I will describe Nick Bostrom's theory of self-locating credences. I will argue that Bostrom's theory of self-locating credences is inadequate. My main argument for this claim will be that Bostrom's theory does not tell us how reference classes should be delineated, and that it is therefore not a complete theory. Third, I will propose a reference class definition in order to develop Bostrom's framework into a complete theory. I will show that this reference class definition makes intuitively correct predictions for the various thought-experiments and illuminates the correct way to think about self-locating reasoning. In the later part of the essay, I will apply my reference class definition to the Sleeping Beauty case, and defend my theory from potential objections.

We often think of non-self-locating propositions as sets of possible worlds. In this essay, we will think of possible worlds as being inhabited by finitely many observers. Although it is also possible to talk about worlds with infinitely many observers, such worlds are beyond the scope of this essay and would not make the discussion much more interesting. We will also think of observers as being composed of observer-moments, these being time-segments of observers. We will think of self-locating propositions as sets of observer-moments. We will also think of non-self-locating propositions as a special case of self-locating proposition: namely, we will consider a non-self-locating proposition *H* to be a set of observer-moments such that for any world *w*, either all observer-moments that are in *w* are in *H* or none of the observer-moments that are in *w* are in *H*. With this background, we are ready to begin describing Arntzenius and Dorr's theory.

Arntzenius and Dorr's Theory

In this section, I will discuss Arntzenius and Dorr's theory. First, I will state the theory itself. Second, I will apply the theory to some thought-experiments to show what it implies about those thought-experiments.

Arntzenius and Dorr's theory of self-locating credences consists of two principles: *Proportion* and the *Principal Principle*, which are stated below.

Proportion: If C is a prior credence function, *H* is a non-self-locating proposition that implies the total duration of the lives of all observers is positive and finite and *F* is any qualitative property, $\langle F : G \rangle$ is the random variable that takes the value of the ratio of the total time, for all things that are ever *F*, during which they are *F* to the total time, for all things that are ever *G*, during which they are *G* and Ĉ(*R*|*H*) is the expected value of variable *R* in Ĉ(.|*H*), then the following equation holds:

$$\mathrm{C}(I\ am\ F|H) = \hat{\mathrm{C}}(\langle F \text{ observer} : \text{observer}\rangle|H)$$[2]

[2] Dorr and Arntzenius, 'Self-locating Priors and Cosmological Measures', p. 12.

In words, this means that a rational observer's prior credence that it is *F* given some hypothesis *H* must be equal to the expected value of the ratio of overall observer-time that is *F* to the overall total observer-time.

Principal Principle: If C is a prior credence function and *A* is a non-self-locating proposition, then $\mathrm{C}(A|P \text{ is the objective chance function}) = P(A)$.[3]

This is the version of the Principal Principle that Arntzenius and Dorr use. The Principal Principle was first formulated by David Lewis. Lewis' version of the principle is notably different from the one Arntzenius and Dorr use, however. In Lewis' version of the Principal Principle, the objective chance function P_{tw} explicitly depends on the time *t* and possible world *w*.[4] Part of the reason for this is that the chance function evolves in such a way that the chance of past events becomes 0 or 1. For example, suppose that a coin is going to be tossed five minutes from now. Up until the coin is tossed, we may suppose that the objective chance of the coin landing heads is 1/2. However, if *t* is a time after the coin toss occurred and *w* is any world in which the coin toss occurred, then P_{tw}(*heads*) will be either 0 or 1. However, Arntzenius and Dorr seem to want the chance function *P* to also constrain an observer's prior credence in past events. Hence, for Arntzenius and Dorr, *P* must be an objective probability function that assigns values other than 0 and 1 even to past events.

What does Arntzenius and Dorr's theory imply about the Incubator thought-experiment? Let *P* denote the objective chance function, let *heads* and *tails* denote the propositions that the outcome of the coin toss is heads or tails respectively, and let C denote the observers' prior credence function. Since the coin is said to be fair, we know that *P*(*heads*) = *P*(*tails*) = 1/2. Moreover, since the only two relevant outcomes in the thought-experiment are *heads* and *tails*, the fact that *P*(*heads*) = *P*(*tails*) = 1/2 tells us all we need to know about

[3] Dorr and Arntzenius, 'Self-locating Priors and Cosmological Measures', p. 15.
[4] Lewis, p. 277.

the objective chance function *P* for the purposes of the thought-experiment. Since all observers know the setup and the only relevant facts about *P* are that *P*(*heads*) = *P*(*tails*) = 1/2, we may say that C(*P* is the objective chance function) = 1. By the Principal Principle, it then follows that C(*heads*) = 1/2, as we can see from the following calculation:

$$\mathrm{C}(heads) = \mathrm{C}(heads|P \text{ is the objective chance function}) = \frac{1}{2}$$

By Proportion, it then follows that C(*heads*|I am in cell #1) = 1/101, as we can see from the following calculation:

$$\mathrm{C}(heads|\text{I am in cell \#1}) = \frac{\mathrm{C}(\text{I am in cell \#1}|heads)\mathrm{C}(heads)}{\mathrm{C}(\text{I am in cell \#1})}$$

$$= \frac{\mathrm{C}(\text{I am in cell \#1}|heads)\mathrm{C}(heads)}{\mathrm{C}(\text{I am in cell \#1}|heads)\mathrm{C}(heads) + \mathrm{C}(\text{I am in cell \#1}|tails)\mathrm{C}(tails)}$$

$$= \frac{\hat{\mathrm{C}}(\langle \text{observers in cell \#1} : \text{observers}\rangle|heads)\mathrm{C}(heads)}{\hat{\mathrm{C}}(\langle \text{observers in cell \#1} : \text{observers}\rangle|heads)\mathrm{C}(heads) + \hat{\mathrm{C}}(\langle \text{observers in cell \#1} : \text{observers}\rangle|tails)\mathrm{C}(tails)}$$

$$= \frac{\frac{1}{100} \times \frac{1}{2}}{\frac{1}{100} \times \frac{1}{2} + 1 \times \frac{1}{2}} = \frac{1}{101}$$

A notable feature of Proportion is that the total number of observers in the universe affects a rational observer's credences in a way which we intuitively may find incorrect. Artnzenius and Dorr use the following thought-experiment to illustrate this:

> **Modified Incubator:** Everything is the same as in the original Incubator thought-experiment, but there are an additional 1 million observers on a distant planet whose experiences are completely different from those of the observers in the incubator and who know the setup. What should be the credence in heads of an observer in the incubator? And what should be the credence in heads of an observer who has exited their cell and observed that they were in cell #1?[5]

[5] Dorr and Arntzenius 'What to Expect in an Infinite World', p. 9.

Let us now consider how Arntzenius and Dorr's theory handles this case. Again, let *P* denote the objective chance function. As in the original Incubator, *P*(*heads*) = *P*(*tails*) = 1/2 since the coin is fair and we can set C(*P* is the objective chance function) = 1 since all observers know the setup and since the only relevant fact about *P* is that *P*(*heads*) = *P*(*tails*) = 1/2. Then, as in the original Incubator, C(*heads*) = 1/2. However, in the Modified Incubator, we obtain the result that $\mathrm{C}(heads|\text{I am in cell \#1}) \cong \frac{1}{2}$, as we can see from the following calculation:

$$\mathrm{C}(heads|\text{I am in cell \#1}) = \frac{\mathrm{C}(\text{I am in cell \#1}|heads)\mathrm{C}(heads)}{\mathrm{C}(\text{I am in cell \#1})}$$

$$= \frac{\mathrm{C}(\text{I am in cell \#1}|heads)\mathrm{C}(heads)}{\mathrm{C}(\text{I am in cell \#1}|heads)\mathrm{C}(heads) + \mathrm{C}(\text{I am in cell \#1}|tails)\mathrm{C}(tails)}$$

$$= \frac{\frac{1}{1{,}000{,}100} \times \frac{1}{2}}{\frac{1}{1{,}000{,}100} \times \frac{1}{2} + \frac{1}{1{,}000{,}001} \times \frac{1}{2}} = \frac{\frac{1}{1{,}000{,}100}}{\frac{2{,}000{,}101}{1{,}000{,}100 \times 1{,}000{,}001}} = \frac{1{,}000{,}001}{2{,}000{,}101} \cong \frac{1}{2}$$

The reason we obtain this result is because we also obtain the result that $\mathrm{C}(heads|\text{I am in the incubator}) \cong \frac{1}{100}$ in the Modified Incubator, as we can see from this calculation:

$$\mathrm{C}(heads|\text{I am in the incubator}) = \frac{\mathrm{C}(\text{I am in the incubator}|heads)\mathrm{C}(heads)}{\mathrm{C}(\text{I am in the incubator})}$$

$$= \frac{\mathrm{C}(\text{I am in the incubator}|heads)\mathrm{C}(heads)}{\mathrm{C}(\text{I am in the incubator}|heads)\mathrm{C}(heads) + \mathrm{C}(\text{I am in the incubator}|tails)\mathrm{C}(tails)}$$

$$= \frac{\frac{1}{1{,}000{,}100} \times \frac{1}{2}}{\frac{100}{1{,}000{,}100} \times \frac{1}{2} + \frac{1}{1{,}000{,}001} \times \frac{1}{2}} = \frac{\frac{1}{1{,}000{,}100}}{\frac{100{,}000{,}100 + 1{,}000{,}100}{1{,}000{,}100 \times 1{,}000{,}001}} = \frac{1{,}000{,}001}{101{,}000{,}200} \cong \frac{1}{100}$$

Arntzenius and Dorr endorse these results about the Modified Incubator. Their treatment of the Modified Incubator seems to rest on the following intuition. An observer who wakes up in the incubator can reason as though they in some sense 'could have been' one of the observers on the planet. Then, the fact that they are in the incubator is evidence in favour of heads, as there are 100 times as many observers in the incubator in the heads-world as there are observers in the incubator in the tails-world.

However, I think that these results about the Modified Incubator are problematic for two related reasons. Firstly, as I have said, Proportion when applied to the Modified Incubator seems to tell the observer to reason as though they could have been another observer; this raises the question of what ‘could have been’ means and in what sense, if any, it is true that an observer could have been another observer. However, if ‘could have been’ is to be understood as merely a heuristic for self-locating reasoning and not in any literal sense true, then the fact that Proportion treats the observer as though they could have been someone else is not in itself necessarily a serious problem for Proportion.

However, there is a second and related problem: Proportion says that an observer should reason as though they could have been *any* other observer that possibly exists, and not only that they could have been some of the other observers that possibly exist. In the Modified Incubator, Proportion tells the observer in the incubator that they should reason as though they could have been an observer on the planet. However, even if observers should reason as though they could have been someone else in self-locating reasoning, there is no reason to additionally say that observers should reason as though they could have been *anyone* else.

Arntzenius and Dorr partly address this point in their discussion of what kinds of beings should be counted as observers when applying Proportion. In order to support the intuition that hypothetical possible intelligent, humanoid lizards should be considered observers when applying Proportion, they ask us to consider the following four theories:

> ‘T0: In any given solar system there is a certain tiny chance ε for life to evolve at all, but if observers do come into existence they are likely to be human-like creatures whose DNA uses two base pairs, having five toes on each foot.

> ‘T1: In any given solar system, the chance of human-like life evolving is ε, but the chance of lizard-like life evolving is $10{,}000\varepsilon$.
>
> ‘T2: In any given solar system, the chance of human-like creatures with five toes on each foot evolving is ε, while the chance of human-like creatures with six toes on each foot evolving is $10{,}000\varepsilon$.
>
> ‘T3: In any given solar system, the chance of human-like creatures whose DNA uses only two base pairs is ε, while the chance of human-like creatures whose DNA uses three or more base pairs is $10{,}000\varepsilon$.’[6]

Arntzenius and Dorr argue that our total evidence strongly confirms T0 over T1, T2 and T3. Their argument for this claim goes roughly as follows. Suppose that we have not yet discovered that human DNA uses two base pairs and that we are considering hypotheses T0 and T3. Whatever our prior credences in T0 and T3 are, the discovery that our DNA uses two base pairs should raise our credence in T0 and lower our credence in T3. Now, suppose that we are considering T0 and T2. The impact of our total evidence on T2 should be approximately the same as it is on T3, because T2 and T3 each say that the chance of human-like creatures that are similar to us, yet different, evolving is 10,000 times greater than the chance of humans evolving. Whether these human-like creatures are different because their DNA uses three base pairs or because they have six toes on each foot does not seem to be a difference that is relevant to self-locating reasoning. Finally, Arntzenius and Dorr challenge us to find a reason not to place the intelligent, lizard-like observers on the same level as human observers:

> ‘Is it their scaly skin? Their cold blood? Their bizarre social structures? Their alien sensory experiences? We have the feeling that as the scenario is fleshed out in more

[6] Dorr and Arntzenius, ‘Self-locating Priors and Cosmological Measures’, p. 19.

> detail, the sense that there could be any important difference between T1 and T2 will start to fade away.'[7]

Hence, Arntzenius and Dorr absorb T1 into T2 and T3. Hence, they argue, our total evidence really does support T0 and go against T1—we should reason as though, a priori, we might have been the hypothetical lizards, despite our great differences from them.

Let us now return to the problems that Modified Incubator poses for Proportion. I have said that although it might make sense for an observer to reason as though they in some sense might have been another observer (even if this is only a heuristic and is not in any literal sense true), the challenge remains for Arntzenius and Dorr to show that an observer should reason as though they might have been *any* other observer. Do Arntzenius and Dorr's remarks about lizard-like beings, human-like beings with six toes or human-like beings with three or more base pairs in their DNA, help us here?

I think that they do not, for the following reason. The characteristics of observers that Arntzenius and Dorr describe here are intrinsic characteristics of beings. But Modified Incubator does not say anything about the intrinsic characteristics of its observers; we can assume that all of the observers in the thought-experiment are human. So, although Arntzenius and Dorr are persuasive in showing that the intrinsic characteristics of observers are irrelevant in self-locating reasoning, they do not show that every observer should always reason as though they might have been any possible observer, because from the point of view of a particular observer, non-intrinsic features of those other observers may make those latter observers irrelevant.

To see this, consider the Modified Incubator thought-experiment again, this time from the point of view of an observer on the planet. We can add the stipulation that the observers on the planet exist before the coin toss occurs and are therefore causally unaffected by the

[7] Dorr and Arntzenius, 'Self-locating Priors and Cosmological Measures', p. 20.

outcome of the coin toss, and that they know this. Now, what should the value of C(*heads*|I am on the planet) be? Intuitively, it should be 1/2. If I know that a fair coin is going to be tossed in the future and that any event that I can observe now is therefore causally unaffected by the outcome of the coin toss, then it is difficult to see how my credence in heads conditional on my total evidence could be anything other than 1/2, regardless of what my total evidence may be.

However, Proportion and the Principal Principle together imply that $\mathrm{C}(heads|\text{I am on the planet}) \neq \frac{1}{2}$, as we can see from the following calculation:

$$\mathrm{C}(heads|\text{I am on the planet}) = \frac{\mathrm{C}(\text{I am on the planet}|heads)\mathrm{C}(heads)}{\mathrm{C}(\text{I am on the planet})}$$

$$= \frac{\mathrm{C}(\text{I am on the planet}|heads)\mathrm{C}(heads)}{\mathrm{C}(\text{I am on the planet}|heads)\mathrm{C}(heads) + \mathrm{C}(\text{I am on the planet}|tails)\mathrm{C}(tails)}$$

$$= \frac{\frac{1{,}000{,}000}{1{,}000{,}100} \times \frac{1}{2}}{\frac{1{,}000{,}000}{1{,}000{,}100} \times \frac{1}{2} + \frac{1{,}000{,}000}{1{,}000{,}001} \times \frac{1}{2}} = \frac{\frac{1}{1{,}000{,}100}}{\frac{1}{1{,}000{,}100} + \frac{1}{1{,}000{,}001}} = \frac{1{,}000{,}001}{2{,}000{,}101} \neq \frac{1}{2}$$

I think that this result, which Arntzenius and Dorr endorse, is undesirable. In the next section, we will explore problems in this vein for Proportion, which will strengthen the intuition that Proportion must be wrong.

Let us now discuss the importance of the notion of observerhood for Proportion. Obviously, applying Proportion requires what counts as an observer to be clearly defined, because an observer's credences depend on the total number of observers in the world. Therefore, if we are to apply Proportion, then we have to know what kinds of beings are observers. Arntzenius and Dorr think that it is impossible to draw a principled and sharp distinction between observers and non-observers.[8] We can show that a clear definition of

[8] Dorr and Arntzenius, 'Self-locating Priors and Cosmological Measures', p. 21.

observerhood is important for Arntzenius and Dorr's theory by considering what happens if there are two definitions of an observer.

So, suppose that there are two different definitions of observer, observer$_1$ and observer$_2$. For example, we may have a world with finitely many chimpanzees and humans (and non-zero of each), some of whom are comatose, so that conscious chimpanzees and non-comatose humans may count as observers$_1$, while comatose and non-comatose humans count as observers$_2$. The classes of observers$_1$ and observers$_2$ are therefore overlapping and there are finitely many observers$_1$ and observers$_2$ in the world. The two definitions of observerhood give us two versions of Proportion, which we may call Proportion$_1$ and Proportion$_2$, and two corresponding prior credence functions, C_1 and C_2 respectively. We may define C_2 in terms of C_1 as follows (*F* stands for any qualitative property):

$$\mathrm{C}_2(I\ am\ F) = \frac{\hat{\mathrm{C}}_1(\langle F \text{ observers}_2\text{: observers}_1\rangle)}{\hat{\mathrm{C}}_1(\langle \text{observers}_2\text{: observers}_1\rangle)}$$

Arntzenius and Dorr show that if we define C_2 this way, then C_2 satisfies Proportion$_2$ and is equivalent to C_1 conditional on 'I am an observer$_1$ and an observer$_2$'.[9]

This result may seem to imply that the question of finding a defined boundary between observers and non-observers in general is not an important problem, as any definition that includes the relevant observers will, in conjunction with its version of Proportion, yield equivalent credence functions conditional on the observer falling under any such definition of an observer. However, this is not the case because of the role played by the Principal Principle in Arntzenius and Dorr's theory. According to Arntzenius and Dorr, the Principal Principle applies to prior credences, and as C_1 and C_2 may be different, in general they will not both satisfy the Principal Principle.[10]

[9] Dorr and Arntzenius, 'What to Expect in an Infinite World', p. 12.
[10] Dorr and Arntzenius, 'What to Expect in an Infinite World', pp. 13.

To illustrate, suppose that in the Modified Incubator example, the observers on a distant planet are comatose humans, so that they count as observers$_2$ but not observers$_1$. Note that all observers$_1$ are observers$_2$ in this scenario, so being an observer$_1$ and an observer$_2$ in this case is equivalent to being an observer$_1$. Let Proportion$_1$ and Proportion$_2$ be the corresponding versions of proportion for the two notions of observerhood. If C_1 and C_2 are prior credence functions that satisfy Proportion$_1$ and Proportion$_2$ respectively and are equivalent conditional on 'I am an observer$_1$' (i.e. they satisfy the equation C_1(H|*I am an observer*$_1$) = C_2(H|*I am an observer*$_1$) for any H), then we can show that they cannot both satisfy the Principal Principle. Suppose that C_2 satisfies the Principal Principle. Then C_2(*heads*) = 1/2. From this and the equation relating C_2 with C_1, it follows that the following equation holds:

$$\mathrm{C}_1(heads) = \frac{\hat{\mathrm{C}}_2(\langle heads\ observers_1 : \ observers_2\rangle)}{\hat{\mathrm{C}}_2(\langle observers_1 : observers_2\rangle)}$$

$$= \frac{\frac{1}{2} \times \frac{100}{1{,}000{,}100} + \frac{1}{2} \times \frac{0}{1{,}000{,}001}}{\frac{1}{2} \times \frac{100}{1{,}000{,}100} + \frac{1}{2} \times \frac{1}{1{,}000{,}001}} = \frac{1{,}000{,}001}{1{,}010{,}002} \cong 0.99$$

Hence, if C_2 satisfies the Principal Principle, then C_1 does not. Since Arntzenius and Dorr's theory uses the Principal Principle to generate prior credences, this shows that the problem of defining what counts as an observer cannot be set aside if we accept Proportion and the Principal Principle.[11]

Arntzenius and Dorr try to resolve this by considering the possibility that an observer can reason as though several different definitions of an observer are possible, without committing to just one definition of observer. One way this might be done is to assign different 'degrees of observerhood' to various kinds of beings, effectively having a version of Proportion that weighs different observers unequally; or, a rational observer's credence function may

[11] Dorr and Arntzenius, 'What to Expect in an Infinite World, p. 14.

range over several different versions of Proportion that use different definitions of observer, instead of assigning a credence of 1 to just one version of Proportion.[12] However, a theory of self-locating credences is more complete the more it tells us about what observers are. Assigning different degrees of observerhood only pushes the question back, as we do not know how we should weigh different kinds of beings with respect to observerhood unless we know what observers are; and if we allow an observer to have various credences in several different versions of Proportion, we still do not know what credence an observer should assign in any particular version of Proportion.

Finally, we should note that a definition of observer is not necessary if an observer's prior credence function is given and we only want to know how an observer should update their credence function in response to a change in their evidential state. I will not repeat Arntzenius and Dorr's proof of this here, but it is easy to intuitively see that it is true. They call this principle *Proportional Update*. If an observer has evidence E and credence function C_t and evidence E^+ and credence function C_{t+} at time t, then Proportional Update states the following:

$$\mathrm{C}_{t+}(I\ am\ F) = \frac{\hat{\mathrm{C}}_t(\langle FE^+\text{observers} : E^+\text{observers}\rangle)}{\hat{\mathrm{C}}_t(\langle E^+\text{ observers} : E\text{ observers}\rangle)}$$ [13]

This equation holds whenever the right-hand side is defined. For this to be the case, the expected ratio of F and E observers to E observers must be finite. It is easy to see that the right-hand side equals the prior credence in F conditional on E^+ divided by prior credence in E^+ conditional on prior credence in E, provided that E^+ entails E. If the denominator equals 1, then $\mathrm{C}_{t+}(I\ am\ F) = \mathrm{C}_t(I\ am\ F | I\ am\ E^+)$.

We have now outlined Arntzenius and Dorr's theory of self-locating credences for situations with finitely many observers. In the next section, I will discuss problems for Anrtzenius and Dorr's theory. I will do this by discussing the Modified Incubator case above, as

[12] Dorr and Arntzenius, 'What to Expect in an Infinite World', p. 15.
[13] Dorr and Arntzenius, 'Self-locating Priors and Cosmological Measures', p. 22.

well as by outlining several thought experiments by Nick Bostrom. These problems motivate Bostrom's own theory, which I will outline in the section after the next section.

Problems for Arntzenius and Dorr

Arntzenius and Dorr's theory has two kinds of features that may be considered problematic. The first feature we have already seen: it is that the total number of observers in a world sometimes affects credences in a way which may seem incorrect. We saw this feature in the two versions of the Incubator case. If the people created by the Incubator are the only observers in the universe, then the credence in heads given the observer is in cell #1 is 1/101, whereas if there are a million additional observers in the world, the posterior credence is close to 1/2. The problem with this result is that the additional one million observers do not seem relevant to the credences of the observer or observers in the incubator. The second problem is illustrated by Bostrom in several thought-experiments which he calls *Adam & Eve*, *UN++* and *Quantum Joe*.

Let us begin by discussing the first problem. Arntzenius and Dorr endorse this as a correct result and do not seek to explain it away. Suppose that you find yourself in the Modified Incubator scenario, but that you have not yet opened your eyes and you therefore do not know whether you are in the incubator or on the planet. By the Principal Principle, your credence in heads is 1/2. What happens when you open your eyes? If you observe that you are on the planet, your credence in heads will not change much and will stay close to 1/2 because your prior credence in being on the planet given heads is at 1,000,000/1,000,100 very close to your prior credence in being on the planet given tails at 1,000,000/1,000,001. If you observe that you are in the incubator, your credence in heads will dramatically increase, because your prior credence that you are in the incubator given heads is almost 100 times larger than your prior credence that you are incubator given tails.

However, this raises the question: why should the prior credence function be equal to the objective chance function? Why should we not instead set the credence in heads conditional on being in the incubator to be equal to 1/2?

Consider the Modified Incubator case again. If we stipulate that the observers on the planet exist before the coin toss occurs and know the setup, including the fact that the coin is fair, then the most reasonable credence in heads conditional on being on the planet is 1/2. If we then also assume Proportion, we obtain the result that C(*heads*|*incubator*) = 100/101, as we can see from the following calculation:

$$\mathrm{C}(heads|\text{I am on the planet}) = \frac{1}{2} = \frac{\mathrm{C}(\text{I am on the planet}|heads)\mathrm{C}(heads)}{\mathrm{C}(\text{I am on the planet})}$$

$$\frac{\mathrm{C}(\text{I am on the planet})}{\mathrm{C}(\text{I am on the planet}|heads)} = 2\mathrm{C}(heads)$$

$$\mathrm{C}(\text{I am on the planet}|heads)\mathrm{C}(heads) + \mathrm{C}(\text{I am on the planet}|tails)\mathrm{C}(tails)$$

$$= 2\mathrm{C}(heads)\mathrm{C}(\text{I am on the planet}|heads)$$

$$\mathrm{C}(\text{I am on the planet}|tails)\big(1 - \mathrm{C}(heads)\big) = \mathrm{C}(heads)\mathrm{C}(\text{I am on the planet}|heads)$$

$$\frac{1{,}000{,}000}{1{,}000{,}001}\big(1 - \mathrm{C}(heads)\big) = \mathrm{C}(heads)\frac{1{,}000{,}000}{1{,}000{,}100}$$

$$\mathrm{C}(heads) = \frac{1{,}000{,}100}{2{,}000{,}101}$$

$$\mathrm{C}(heads|\text{I am in the incubator}) = \frac{\mathrm{C}(\text{I am in the incubator}|heads)\mathrm{C}(heads)}{\mathrm{C}(\text{I am in the incubator})}$$

$$= \frac{\mathrm{C}(\text{I am in the incubator}|heads)\mathrm{C}(heads)}{\mathrm{C}(\text{I am in the incubator}|heads)\mathrm{C}(heads) + \mathrm{C}(\text{I am in the incubator}|tails)\mathrm{C}(tails)}$$

$$= \frac{\frac{100}{1{,}000{,}100} \times \frac{1{,}000{,}100}{2{,}000{,}101}}{\frac{100}{1{,}000{,}100} \times \frac{1{,}000{,}100}{2{,}000{,}101} + \frac{1}{1{,}000{,}001} \times \frac{1{,}000{,}001}{2{,}000{,}101}} = \frac{100}{101}$$

I think that this value for C(*heads*|*incubator*) is undesirable for the following reason. We want the value of C(*heads*|*incubator*) to be the same regardless of how many additional observers there are aside from those in the incubator, because it is not in any sense true that an observer in the incubator could have been an observer on the planet. Rather, if we allow that observers can be identified by whether they are in the heads-world or in the tails-world and what cell they are in (and I think that this is an unproblematic assumption), then an observer in the incubator is in a state of not knowing which observer they are, and given that they are in the cell, their evidence cannot shift their credence in heads away from 1/2. Then, in Modified Incubator, C(*heads*|I am in the incubator) should be equal to 1/2, as in the original Incubator.

To sum up, we want C(*heads*|I am on the planet) = C(*heads*|I am in the incubator) = 1/2 to hold. However, given Proportion and C(*heads*|I am on the planet) = 1/2, we obtain the result C(*heads*|I am in the incubator) = 100/101. If we want C(*heads*|I am in the incubator) = 1/2, we have to reject Proportion.

Let us now turn to the second problematic feature of Arntzenius and Dorr's theory. Nick Bostrom illustrates this problem with a series of thought-experiments which he calls Adam & Eve, UN++ and Quantum Joe. These thought experiments seem to suggest that Arntzenius and Dorr's theory implies that observers can have powers such as anomalous precognition, psychokinesis and backward causation. Let us now outline these thought experiments.

Adam & Eve: Adam and Eve want to avoid conceiving a child. If they conceive a child, then they will likely be the first of hundreds of billions of observers in the world, and if they do not conceive a child, then they will be the only two observers in the world. In this case, a rational observer's credence that they are Adam or Eve conditional on the hypothesis that conception occurs is very low, while their credence that they are Adam or Eve conditional on the hypothesis that conception does not occur is 1. Therefore, if we accept Arntzenius and Dorr's theory, Adam's and Eve's credence that conception will occur (conditional on them being Adam or Eve) should be very low and much lower than the objective chance of conception. This seems to show that Adam and Eve have a kind of anomalous precognition.[14]

UN++: In the year 2100, the world is ruled by an all-powerful world government, the UN++. Due to technological advances and the UN++'s power, any decision made by the UN++ is certain to be enacted. Signs have been detected that there will be a series of *n* gamma ray bursts. For each burst in the series, there is a 90% objective chance that it will happen. In order to prevent the total number of humans in the universe decreasing as a result of gamma ray bursts, the UN++ makes the following decision: for every gamma ray burst that happens, the UN++ will build enough space colonies such that the number of humans alive is increased by a factor of *m* compared to the number alive before the gamma ray burst. If *m* is sufficiently great compared to *n*, the *UN++* can now have high credence that no gamma ray bursts will happen. This is obtained by similar reasoning as in the Adam & Eve thought experiments (namely, that humans alive before any single gamma ray burst are unlikely to be very early in the birth order of all humans who will ever have lived). The crucial difference between this thought-experiment and the *Adam & Eve* thought experiments is that the hypothetical UN++ exists in the future from us, whereas early humans existed in our past. Since the UN++ possibly exists in the future, this seems to suggest that we have the ability to form the intention to create the UN++, which would apparently have the power to stop gamma ray bursts. But this would mean that we have the power to stop gamma ray bursts. Bostrom says that we clearly do not have the ability to prevent future gamma ray bursts, and that we should therefore reject what he calls the 'self-sampling assumption' (of which Proportion is a version), which is used in reaching the conclusion that we have this ability.[15]

[14] Bostrom, p. 142.
[15] Bostrom, pp. 150–151.

Quantum Joe: The scientist Joe has discovered that he is the only observer in the universe so far. He has built a quantum machine that has a 1/10 chance of outputting any one-digit integer, and a reproduction machine which, when activated, will create 10,000 additional observers. Joe then connects the two machines so that any number except zero on the quantum machine will activate the reproduction machine, while zero will cause the reproduction machine to be destroyed. There are not enough materials to build another reproduction machine, so if the reproduction machine is destroyed, then Joe will have been the only observer in the universe. If Joe is about to press the button on the quantum machine, what should his credence in zero be?[16] By the Principal Principle, his credence in zero should be 1/10. However, Arntzenius and Dorr's theory contradicts this given some assumptions. First, assume that Joe's prior credence in zero must be 1/10. Second, let us look at Joe's prior credence that he is Joe ('*zero*' denotes the proposition that the outcome of the quantum machine will be zero):

$$\mathrm{C}(I\ am\ Joe) = \hat{\mathrm{C}}(\langle Joe : observers\rangle) = \mathrm{C}(zero) \times \frac{1}{10{,}001} + \big(1 - \mathrm{C}(zero)\big) \times 1$$

$$= \frac{1}{100{,}010} + \frac{9}{10} = \frac{9{,}001}{10{,}001} \cong \frac{9}{10}$$

Then, Joe's credence in zero given that he is Joe must be the following:

$$\mathrm{C}(zero|I\ am\ Joe) = \frac{\mathrm{C}(I\ am\ Joe|zero)\mathrm{C}(zero)}{\mathrm{C}(I\ am\ Joe)} = \frac{\hat{\mathrm{C}}(\langle Joe : observers\rangle|zero)\frac{1}{10}}{\frac{9{,}001}{10{,}001}}$$

$$= \frac{\frac{1}{10{,}001} \times \frac{1}{10}}{\frac{9{,}001}{10{,}001}} = \frac{1}{90{,}010}$$

Given that Joe's existence and qualitative properties will not in any way be directly causally affected by the result on the quantum machine, this value for Joe's posterior credence in zero

[16] Bostrom, p. 154.

goes strongly against common sense. The common-sense view would be that Joe's posterior credence in zero should be equal to the objective probability of zero, which is 1/10. As in the case of the modified Incubator, we would obtain the common-sense result by applying the Principal Principle not to an observer's prior credence function but to their (Joe's in this case) posterior credence function conditional on their total evidence.

This concludes our discussion of the problems for Arntzenius and Dorr's theory. In the next section, I will outline Bostrom's theory, which attempts to solve these problems. In the section after the next section, I will outline problems for Bostrom's theory.

Bostrom's Theory

Now, let us look at Bostrom's theory. The crucial difference between Arntzenius and Dorr's theory and Bostrom's is that while the former puts all observer-moments on equal footing, the latter employs the notion of a reference class, which is a subset of all observer-moments in a range of possible worlds. We will discuss the notion of a reference class in more detail later. For now, let us state Bostrom's equation.

Let α denote an observer-moment, let P_α denote α's prior credence function, let w_α denote the world that α is located in, let Ω_α denote the set of observer-moments in the same reference class as α, let $\Omega(w)$ denote the set of observer-moments in world w and let Ω_h denote the set of observer-moments of whom h is true (if h is a self-locating proposition, this means h observer-moments; if h is non-indexical, it means observer-moments in worlds where h is true). Bostrom proposes the following equation that a rational observer's credence function must obey the so-called *Observation Equation*:

$$P_\alpha(h|e) = \frac{1}{\gamma} \sum_{\sigma \in \Omega_h \cap \Omega_e} \frac{P_\alpha(w_\sigma)}{|\Omega_\sigma \cap \Omega(w_\sigma)|}, \gamma = \sum_{\sigma \in \Omega_e} \frac{P_\alpha(w_\sigma)}{|\Omega_\sigma \cap \Omega(w_\sigma)|}$$ [17]

[17] Bostrom, pp. 172–173.

We can best understand Bostrom's overall approach if we contrast it with that of Arntzenius and Dorr. In Arntzenius and Dorr's framework, there are possible worlds and observer-moments that inhabit those possible worlds, as well as an objective chance function that ranges over the set of worlds. Strictly speaking, Arntzenius and Dorr do not explicitly refer to possible worlds, but only talk about self-locating and qualitative propositions. However, nothing essential is changed if we think in terms of possible worlds inhabited by observer-moments, with propositions as sets of observer-moments.

Bostrom's framework has two key differences. First, associated with every observer-moment is a reference class, which is a set of observer-moments. Although Bostrom never explicitly says this, it is clear from the examples he uses that the set of reference classes is a partition of the set of all observer-moments. This means that every observer-moment belongs to exactly one reference class. Secondly, Bostrom's theory does not tell us what the prior credence should be in any possible world. It only tells us what the credence should be in being a particular observer-moment if credences in possible worlds are already given. This is in contrast with Arntzenius and Dorr's theory, which contains the Principal Principle, which says that the prior credence in any given world should be equal to the objective probability of that world being actual.

Furthermore, it is notable that the Observation Equation is stated as a conditional credence function (i.e. we have $P_\alpha(h|e)$ on the left-hand side of the equation rather than, say, $P_\alpha(h)$). This is because the Observation Equation gives results that make sense only if we are evaluating credences for an observer-moment whose total evidential state is consistent with only one reference class. In other words, if we want to use the Observation Equation to evaluate $P_\alpha(h|e)$, then the result will make sense only if *e* rules out all reference classes except one. Intuitively, *e* is meant to represent an observer-moment's total evidential state. Although Bostrom never explicitly says so, it is clear from his discussion that he assumes that an observer-moment's total evidential state alone has to be enough to determine which reference class they belong to; in other words, that two observer-moments cannot have the same total evidential state while being in different reference classes.

Now, let us consider the Observation Equation again in order to clarify what it says. Let Ω_1 denote some reference class, and let e denote the proposition that I (the observer) belong in Ω_1. Then, $\Omega_e = \Omega_1$, since Ω_e is the set of e observers, which is the reference class Ω_1. Also, let W_e denote the set of all possible worlds that contain at least one observer in reference class Ω_1, that is, the set of all worlds w such that $\Omega_1 \cap \Omega(w) \neq \emptyset$. If we insert this into the Observation Equation and h denotes any proposition, we get the following result:

$$\gamma = \sum_{\sigma \in \Omega_e} \frac{P_\alpha(w_\sigma)}{|\Omega_\sigma \cap \Omega(w_\sigma)|} = \sum_{w \in W_e} \sum_{\sigma \in \Omega_1 \cap \Omega(w)} \frac{P_\alpha(w)}{|\Omega_1 \cap \Omega(w)|} = \sum_{w \in W_e} P_\alpha(w)$$

$$P_\alpha(h|e) = \frac{1}{\gamma} \sum_{\sigma \in \Omega_h \cap \Omega_e} \frac{P_\alpha(w_\sigma)}{|\Omega_\sigma \cap \Omega(w_\sigma)|} = \frac{1}{\gamma} \sum_{\sigma \in \Omega_h \cap \Omega_1} \frac{P_\alpha(w_\sigma)}{|\Omega_1 \cap \Omega(w_\sigma)|}$$

$$= \frac{1}{\gamma} \sum_{w \in W_e} \sum_{\Omega_h \cap \Omega_1 \cap \Omega(w)} \frac{P_\alpha(w)}{|\Omega_1 \cap \Omega(w)|} = \frac{1}{\gamma} \sum_{w \in W_e} \frac{|\Omega_h \cap \Omega_1 \cap \Omega(w)| P_\alpha(w)}{|\Omega_1 \cap \Omega(w)|}$$

$$= \frac{\sum_{w \in W_e} \frac{|\Omega_h \cap \Omega_1 \cap \Omega(w)| P_\alpha(w)}{|\Omega_1 \cap \Omega(w)|}}{\sum_{w \in W_e} P_\alpha(w)}$$

The normalisation constant γ is equal to the probability that at least one observer-moment in the reference class Ω_1 exists. The numerator of the above fraction is the expectation of the ratio of the number of observer-moments in Ω_1 that are h to the total number of observer-moments in Ω_1.

Now, if we let h be the proposition that a world w_1 is actual, we get the following result:

$$P_\alpha(h|e) = \frac{\sum_{w \in W_e} \frac{|\Omega_h \cap \Omega_1 \cap \Omega(w)| P_\alpha(w)}{|\Omega_1 \cap \Omega(w)|}}{\sum_{w \in W_e} P_\alpha(w)} = \frac{\sum_{w \in W_e} \frac{|\Omega(w_1) \cap \Omega_1 \cap \Omega(w)| P_\alpha(w)}{|\Omega_1 \cap \Omega(w)|}}{\sum_{w \in W_e} P_\alpha(w)}$$

$$= \frac{\frac{|\Omega_1 \cap \Omega(w_1)| P_\alpha(w_1)}{|\Omega_1 \cap \Omega(w_1)|}}{\sum_{w \in W_e} P_\alpha(w)} = \frac{P_\alpha(w_1)}{\sum_{w \in W_e} P_\alpha(w)}$$

Hence, the credence in a particular world conditional on being in a particular reference class is equal to the credence in that world divided by the credence that an observer-moment in that reference class exists.

Lastly, let σ_1 denote a particular observer-moment in world w_1 and reference class Ω_1. If we let let h be the proposition that I (this observer-moment) am σ_1, we get the following result:

$$P_\alpha(h|e) = \frac{\sum_{w \in W_e} \frac{|\Omega_h \cap \Omega_1 \cap \Omega(w)| P_\alpha(w)}{|\Omega_1 \cap \Omega(w)|}}{\sum_{w \in W_e} P_\alpha(w)} = \frac{\frac{P_\alpha(w_1)}{|\Omega_1 \cap \Omega(w_1)|}}{\sum_{w \in W_e} P_\alpha(w)}$$

This shows that given that one is in a particular world and reference class, the probability of being any particular observer-moment is equal to 1 divided by the number of observer-moments in that world and reference class, multiplied by the credence in that world, divided by the probability that that reference class exists.

Hence, what the Observation Equation does is the following: given a reference class, it first assigns the credence 0 to any worlds where the reference class exists and assigns a credence proportional to the prior credence in each world where the reference class does exist; then, within each world it assigns the credence 0 to any observer-moment outside of the reference class, and assigns the same credence to each observer-moment inside the reference class.

This shows that the Observation Equation is very similar to Arntzenius and Dorr's Proportion, which sets the prior credence of every world equal to its objective probability and then assigns the same credence to every observer-moment within any given world. Indeed, the Observation Equation is equivalent to Proportion if we stipulate that there is only one reference class.

Now, let us apply Bostrom's theory to the various thought-experiments that we have mentioned above, and show how reference classes can be used in order to avoid the undesirable results that Proportion and the Principal Principle imply.

First, let us consider the original Incubator. We have two worlds, the heads-world, which we may call w_1 and the tails-world, which we may call w_2. It will be adequate to treat every observer as one observer-moment. Hence, there are 100 observer-moments in $\Omega(w_1)$, the set of observer-moments in the heads-world, and 1 observer-moment in $\Omega(w_2)$. We can assign all observer-moments to the same reference class because the Observation Equation is equivalent to Proportion if we assume that there is only one reference class, and Proportion gives the intuitively correct result for the original Incubator. Also, we can set the prior credence in the heads-world and in the tails-world to be equal to 1/2, since the Principal Principle together with Proportion gives the intuitively correct result for the original Incubator. Then, we obtain the following result for $P_\alpha(heads|incubator)$ ('incubator' refers to any observer-moment that is initially in the incubator, so it refers to all observer-moments):

$$P_\alpha(h|e) = \frac{\sum_{\sigma\in\Omega_h\cap\Omega_e}\frac{P_\alpha(w_\sigma)}{|\Omega_\sigma\cap\Omega(w_\sigma)|}}{\sum_{\sigma\in\Omega_e}\frac{P_\alpha(w_\sigma)}{|\Omega_\sigma\cap\Omega(w_\sigma)|}}$$

$$P_\alpha(heads|incubator) = \frac{\sum_{\sigma\in\Omega_{heads}\cap\Omega_{incubator}}\frac{P_\alpha(w_\sigma)}{|\Omega_\sigma\cap\Omega(w_\sigma)|}}{\sum_{\sigma\in\Omega_{incubator}}\frac{P_\alpha(w_\sigma)}{|\Omega_\sigma\cap\Omega(w_\sigma)|}} = \frac{\sum_{\sigma\in\Omega_{heads}\cap\Omega_{incubator}}\frac{P_\alpha(w_1)}{|\Omega(w_1)|}}{\sum_{\sigma\in\Omega_{incubator}}\frac{P_\alpha(w_\sigma)}{|\Omega(w_\sigma)|}}$$

$$= \frac{\frac{100P_\alpha(w_1)}{100}}{\frac{100P_\alpha(w_1)}{100}+\frac{1P_\alpha(w_2)}{1}} = \frac{P_\alpha(w_1)}{P_\alpha(w_1)+P_\alpha(w_2)} = \frac{\frac{1}{2}}{\frac{1}{2}+\frac{1}{2}} = \frac{1}{2}$$

Next, we should consider what happens when we conditionalise one the evidence that the observer-moment is in cell #1. Note that it is adequate to treat every observer as a single observer-moment in this scenario despite the fact that the observers' evidential states change as they exit the cell. We are therefore not complicating this example by treating every observer, before and after they exit their cell, as a single observer-moment. Then, we obtain the following result for $P_\alpha(heads|cell\ \#1)$:

$$P_\alpha(heads|cell\ \#1) = \frac{\sum_{\sigma\in\Omega_{heads}\cap\Omega_{cell\ \#1}} \frac{P_\alpha(w_\sigma)}{|\Omega_\sigma \cap \Omega(w_\sigma)|}}{\sum_{\sigma\in\Omega_{cell\ \#1}} \frac{P_\alpha(w_\sigma)}{|\Omega_\sigma \cap \Omega(w_\sigma)|}} = \frac{\sum_{\sigma\in\Omega_{heads}\cap\Omega_{cell\ \#1}} \frac{P_\alpha(w_1)}{|\Omega(w_1)|}}{\sum_{\sigma\in\Omega_{cell\ \#1}} \frac{P_\alpha(w_\sigma)}{|\Omega(w_\sigma)|}}$$

$$= \frac{\frac{1P_\alpha(w_1)}{100}}{\frac{1P_\alpha(w_1)}{100} + \frac{1P_\alpha(w_2)}{1}} = \frac{\frac{1}{100}\times\frac{1}{2}}{\frac{1}{100}\times\frac{1}{2}+\frac{1}{2}} = \frac{\frac{1}{100}}{\frac{101}{100}} = \frac{1}{101}$$

This is the same result that Arntzenius and Dorr's theory gives, and it is the desired result.

Now, let us consider the Modified Incubator thought-experiment. Recall that in the Modified Incubator, there are 1 million observers on the planet and either 100 or 1 observers in the incubator depending on the outcome of the coin toss. Heads creates one observer in each of 100 cells while tails creates one observer in cell #1. It is adequate to treat each observer as one observer-moment, despite the fact that the total evidential state of the observers in the incubator changes when they exit the incubator. Again, we can suppose that there are two worlds, w_1 and w_2, where w_1 is the heads-world while w_2 is the tails-world. There are 1,000,000 + 100 = 1,000,100 observer-moments in $\Omega(w_1)$, the set of observer-moments in w_1, and 1,000,000 + 1 = 1,000,001 observer-moments in $\Omega(w_2)$. Recall that in the Modified Incubator, the desired credence function is one that assigns a credence of 1/2 to heads conditional on being on the planet and also assigns a credence of 1/2 in heads conditional on being in the incubator, and a credence of 1/101 in heads conditional on being in cell #1. In order to obtain the desired results, we have to let $P_\alpha(w_1) = P_\alpha(w_2) = 1/2$ and we have to place the observer-moments on the planet into one reference class, which we may call Ω_1, and the observer-moments that start out in the incubator into another reference class, which we may call Ω_2. Then, we obtain the following values for sizes of sets of observer-moments:

$$|\Omega_{heads} \cap \Omega_{incubator}| = |\Omega_2 \cap \Omega(w_1)| = 100$$

$$|\Omega_{tails} \cap \Omega_{incubator}| = |\Omega_2 \cap \Omega(w_2)| = 1$$

$$|\Omega_{heads} \cap \Omega_{planet}| = |\Omega_1 \cap \Omega(w_1)| = 1{,}000{,}000$$

$$|\Omega_{tails} \cap \Omega_{planet}| = |\Omega_1 \cap \Omega(w_1)| = 1{,}000{,}000$$

Then, we obtain the following results:

$$P_\alpha(heads|incubator) = \frac{\sum_{\sigma\in\Omega_{heads}\cap\Omega_{incubator}} \frac{P_\alpha(w_\sigma)}{|\Omega_\sigma \cap \Omega(w_\sigma)|}}{\sum_{\sigma\in\Omega_{incubator}} \frac{P_\alpha(w_\sigma)}{|\Omega_\sigma \cap \Omega(w_\sigma)|}}$$

$$= \frac{\sum_{\sigma\in\Omega_{heads}\cap\Omega_{incubator}} \frac{P_\alpha(w_1)}{|\Omega_2 \cap \Omega(w_1)|}}{\sum_{\sigma\in\Omega_{heads}\cap\Omega_{incubator}} \frac{P_\alpha(w_1)}{|\Omega_2 \cap \Omega(w_1)|} + \sum_{\sigma\in\Omega_{tails}\cap\Omega_{incubator}} \frac{P_\alpha(w_2)}{|\Omega_2 \cap \Omega(w_2)|}}$$

$$= \frac{\frac{|\Omega_2 \cap \Omega(w_1)|P_\alpha(w_1)}{|\Omega_2 \cap \Omega(w_1)|}}{\frac{|\Omega_2 \cap \Omega(w_1)|P_\alpha(w_1)}{|\Omega_2 \cap \Omega(w_1)|} + \frac{|\Omega_2 \cap \Omega(w_2)|P_\alpha(w_2)}{|\Omega_2 \cap \Omega(w_2)|}} = \frac{P_\alpha(w_1)}{P_\alpha(w_1) + P_\alpha(w_2)} = \frac{\frac{1}{2}}{\frac{1}{2} + \frac{1}{2}}$$

$$= \frac{1}{2}$$

$$P_\alpha(heads|planet) = \frac{\sum_{\sigma\in\Omega_{heads}\cap\Omega_{planet}} \frac{P_\alpha(w_\sigma)}{|\Omega_\sigma \cap \Omega(w_\sigma)|}}{\sum_{\sigma\in\Omega_{planet}} \frac{P_\alpha(w_\sigma)}{|\Omega_\sigma \cap \Omega(w_\sigma)|}}$$

$$= \frac{\sum_{\sigma\in\Omega_{heads}\cap\Omega_{planet}} \frac{P_\alpha(w_1)}{|\Omega_1 \cap \Omega(w_1)|}}{\sum_{\sigma\in\Omega_{heads}\cap\Omega_{incubator}} \frac{P_\alpha(w_1)}{|\Omega_1 \cap \Omega(w_1)|} + \sum_{\sigma\in\Omega_{tails}\cap\Omega_{incubator}} \frac{P_\alpha(w_2)}{|\Omega_1 \cap \Omega(w_2)|}}$$

$$= \frac{\frac{|\Omega_1 \cap \Omega(w_1)|P_\alpha(w_1)}{|\Omega_1 \cap \Omega(w_1)|}}{\frac{|\Omega_1 \cap \Omega(w_1)|P_\alpha(w_1)}{|\Omega_1 \cap \Omega(w_1)|} + \frac{|\Omega_1 \cap \Omega(w_2)|P_\alpha(w_2)}{|\Omega_1 \cap \Omega(w_2)|}} = \frac{P_\alpha(w_1)}{P_\alpha(w_1) + P_\alpha(w_2)} = \frac{\frac{1}{2}}{\frac{1}{2} + \frac{1}{2}}$$

$$= \frac{1}{2}$$

$$P_\alpha(heads|cell\ \#1) = \frac{\sum_{\sigma\in\Omega_{heads}\cap\Omega_{cell\ \#1}} \frac{P_\alpha(w_\sigma)}{|\Omega_\sigma \cap \Omega(w_\sigma)|}}{\sum_{\sigma\in\Omega_{cell\ \#1}} \frac{P_\alpha(w_\sigma)}{|\Omega_\sigma \cap \Omega(w_\sigma)|}}$$

$$= \frac{\sum_{\sigma\in\Omega_{heads}\cap\Omega_{cell\ \#1}} \frac{P_\alpha(w_1)}{|\Omega_2 \cap \Omega(w_1)|}}{\sum_{\sigma\in\Omega_{heads}\cap\Omega_{cell\ \#1}} \frac{P_\alpha(w_1)}{|\Omega_2 \cap \Omega(w_1)|} + \sum_{\sigma\in\Omega_{tails}\cap\Omega_{cell\ \#1}} \frac{P_\alpha(w_2)}{|\Omega_2 \cap \Omega(w_2)|}}$$

$$= \frac{\frac{1}{100}}{\frac{1}{100} + \frac{1}{1}} = \frac{1}{101}$$

These are the desired results.

Now, I will briefly outline how reference classes can be used to resolve the remaining thought-experiments that Bostrom describes: the three Adam and Eve thought experiments, UN++ and Quantum Joe. I will not set them out in as much detail as the foregoing thought-experiments of Incubator and Modified Incubator because it should be easy to see by analogy with the latter that applying the Observation Equation in the way I describe will give us the desired results.

In the Adam and Eve thought-experiments, we have two worlds: a conception-world and a non-conception world. Both worlds contain Adam and Eve, and the conception-world also contains hundreds of billions of additional observers. If we place Adam and Eve in a separate reference class from their progeny, we will get the desired results: that the probability of conception conditional on being Adam or Eve is equal to the prior probability of conception, while the probability of conception conditional on being one of Adam and Eve's descendants is 1.

In the UN++ thought-experiment, there are *n* possible gamma ray bursts, and for each possible gamma ray burst, the objective probability that it will occur is 9/10. There is a group of observers who are alive in the year 2100 on Earth and who will have existed regardless of which of the possible gamma ray bursts occur. There are also possible groups of observers in space colonies who will exist if some of the possible gamma ray bursts occur. There are 2^n possible worlds, because each of the *n* gamma ray bursts either will or will not occur. The objective probability of each world depends on the number of gamma ray bursts that occur in that world; if *k* bursts occur, then the probability of the world is $0.9^k 0.1^{n-k}$. The observers on Earth exist in all of these possible worlds, while the observers in space colonies exist in only some of the worlds. The desired result is that, for any given possible gamma ray burst, the credence that that burst will occur conditional on being an observer on Earth is 9/10. Since the observers on Earth exist in every possible world and are in a separate reference class from all other possible observers, and since the objective probabilities of all worlds where any particular gamma ray burst occurs will add up to 9/10, it is easy to verify that the credence in any particular gamma ray burst occurring conditional on being an Earth-observer will be 9/10.

Finally, let us consider Quantum Joe. There are ten possible worlds, each corresponding to an integer that the quantum device could output. Joe exists in all of the worlds, while nine of the worlds also contain 10,000 clones of Joe. The objective probability of each world is 1/10. The desired result is that Joe's credence in each world is 1/10. If we place Joe in one reference class and his 10,000 possible clones in another reference class, it is easy to verify that the credence in any particular world conditional on being Joe will be 1/10.

Problems for Bostrom's Theory

In this section, I will discuss the problems and sources of imprecision in Bostrom's theory. The main problem for Bostrom's theory is the fact that it does not contain rules about how reference classes should be delineated. The second problem is that if we were to supplement the Observation Equation with a reference class definition, it seems likely that it would always be possible to come up with thought-experiments in which an observer-moment's evidence is consistent with more than one reference class, which would cause the

Observation Equation to give undesirable results. This second problem is only a potential problem, since it might be possible to find a satisfactory reference class definition that never puts two observer-moments with the same total evidential state into two different reference classes. The third problem is that Bostrom does not tell us how observers should be partitioned into observer-moments. Lastly, the fourth problem is that in the Observation Equation, the credence function P_α is indexed to the observer-moment α, suggesting that different observer-moments should have different prior credence functions. I will now discuss these problems in turn.

The main problem for Bostrom's theory is that he does not give any rules about how observer-moments should be partitioned into reference classes. Instead, Bostrom simply divides observers into reference classes *ad hoc* in such a way that it yields the intuitively correct credence function. Clearly, this is not sufficient for a complete theory of self-locating credences, and Bostrom would have to propose a general reference class definition in order for his theory to be complete.

Bostrom considers two reference class definitions and rejects them. First, he considers R^U, which says that any observer-moment's reference class is the class of all observer-moments. As we have seen, this would make the Observation Equation equivalent to Proportion, so he rejects this reference class definition. Second, he considers the 'minimal' reference class definition R^0, which says that any observer-moment σ's reference class is the class of all observer-moments that are subjectively indistinguishable from the σ—in other words, the set of those observer-moments with the same total evidence. However, he rejects R^0 because the Observation Equation in conjunction with R^0 has the following result: for any observer-moment α whose total evidential state is represented by e and world w_i, if $|\Omega_\alpha \cap \Omega(w_i)| > 0$ then $P_\alpha(w_i|e) = P_\alpha(w_i)$. It is straightforward to show that this is the case if we note that R^0 and $\sigma \in \Omega_e$ implies that $\Omega_\sigma = \Omega_e$:

$$\gamma = \sum_{\sigma \in \Omega_e} \frac{P_\alpha(w_\sigma)}{|\Omega_e \cap \Omega(w_\sigma)|} = \sum_{w \in W} \sum_{\sigma \in \Omega_e \cap \Omega(w)} \frac{P_\alpha(w)}{|\Omega_e \cap \Omega(w)|} = \sum_{w \in W} P_\alpha(w) = 1$$

$$P_\alpha(w_i|e) = \frac{1}{\gamma} \sum_{\sigma \in \Omega_e \cap \Omega(w_i)} \frac{P_\alpha(w_i)}{|\Omega_e \cap \Omega(w_i)|} = P_\alpha(w_i)$$

In words, this means that if at least one observer-moment with the total evidential state *e* exists in w_i then *e* does not confirm w_i. But this is counterintuitive; for example, suppose that a very small proportion of observer-moments in a world w_i have the experience *e* while all observer-moments in another world w_j have the experience *e*. In that case, we would say that *e* confirms w_j, but the reference class definition R^0 implies that it does not.[18]

Bostrom concludes that both R^U and R^0 must be false, and that the correct reference class definition must therefore be an intermediate definition between the two extremes. It is clear from this latter remark that Bostrom assumes that any reference class definition must be coarser-grained than R^0, or, in other words, that if two observers have the same evidence then they must belong to the same reference class on any licit reference class definition. Note that this implies that every observer-moment's total evidential state must be consistent with only one reference class.

A second potential problem for Bostrom's theory is that the Observation Equation gives counterintuitive results when trying to evaluate $P_\alpha(h|e)$ if *e* is consistent with more than one reference class within a single world. Although Bostrom assumes that any licit reference class definition must assign two observer-moments to the same reference class if they are in the same total evidential state, it is possible that if there is a correct reference class definition *R*, that *R* sometimes assigns observer-moments that are in the same total evidential state and the same world to different reference classes. However, in that case the Observation Equation gives counterintuitive results. This can be illustrated with the following thought-experiment:

[18] Bostrom, pp. 177–178.

Modified Adam & Eve: This thought-experiment is the same as the original Adam & Eve, but with the following modification. Every observer, has their memory erased on the midnight before their 20th birthday, and their eyes are kept closed for 24 hours so they cannot know whether they are Adam or Eve or one of their descendants. At midnight after their 20th birthday, their eyes are opened and regain their memories. All observers know the setup. What should an observer's credence that they are Adam or Eve be while their eyes are closed? What should their credence that conception will have occurred be while their eyes are closed?

There is more than one way to represent Modified Adam & Eve in Bostrom's framework. One possibility would be to represent each observer as two observer-moments—a closed-eyed and an open-eyed observer-moment—and to assign all the closed-eyed observer-moments to the same reference class. However, I will not do this now, because what I am now doing is showing that if we represent a thought-experiment in such a way that some total evidential state is consistent with more than one reference class within a single world, the Observation Equation gives counterintuitive results.

For simplicity, we will represent every observer by one observer-moment. Let Ω_1 be the reference class containing Adam and Eve and let Ω_2 be the reference class containing their descendants. Let w_1 be the world in which conception will have occurred and let w_2 be the world in which conception will not have occurred. Then, $|\Omega_1 \cap \Omega(w_1)| = |\Omega_1 \cap \Omega(w_2)| = 2$, $|\Omega_2 \cap \Omega(w_2)| = 0$ and let $|\Omega_1 \cap \Omega(w_1)| = 10^{11}$. A closed-eyed observer lacks any evidence—they could be anyone—so in order to find a closed-eyed observer's credence that they are Adam or Eve, we evaluate $P_\alpha(\Omega_1|\mathrm{T})$:

$$P_\alpha(\Omega_1|\mathrm{T}) = \frac{\sum_{\sigma\in\Omega_1} \frac{P_\alpha(w_\sigma)}{|\Omega_1 \cap \Omega(w_\sigma)|}}{\sum_{\sigma\in\Omega_\mathrm{T}} \frac{P_\alpha(w_\sigma)}{|\Omega_\sigma \cap \Omega(w_\sigma)|}}$$

$$= \frac{\sum_{\sigma\in\Omega_1\cap\Omega(w_1)} \frac{P_\alpha(w_1)}{|\Omega_1 \cap \Omega(w_1)|} + \sum_{\sigma\in\Omega_1\cap\Omega(w_2)} \frac{P_\alpha(w_2)}{|\Omega_1 \cap \Omega(w_2)|}}{\sum_{\sigma\in\Omega_1\cap\Omega(w_1)} \frac{P_\alpha(w_1)}{|\Omega_1 \cap \Omega(w_1)|} + \sum_{\sigma\in\Omega_1\cap\Omega(w_2)} \frac{P_\alpha(w_2)}{|\Omega_1 \cap \Omega(w_2)|} + \sum_{\sigma\in\Omega_2\cap\Omega(w_1)} \frac{P_\alpha(w_1)}{|\Omega_1 \cap \Omega(w_1)|}}$$

$$= \frac{P_\alpha(w_1) + P_\alpha(w_2)}{P_\alpha(w_1) + P_\alpha(w_2) + P_\alpha(w_1)} = \frac{1}{1 + P_\alpha(w_1)}$$

I think that this result is undesirable because it says that a closed-eyed observer's credence that they are Adam or Eve depends only on the probability of conception and not at all on the number of descendants in the world where conception occurs. For example, suppose that the prior probability of conception is $P_\alpha(w_1) = 1/2$. Then, a closed-eyed observer's credence that they are Adam or Eve is $P_\alpha(\Omega_1|\mathrm{T}) = 2/3$. The issue here is not that a credence of 2/3 is too high or too low, or that it does not depend on the number of descendants in the conception-world, but that it is precisely 2/3 regardless of the number of descendants there might be, which seems like an arbitrary value without any justification aside from the fact that that is the value that the Observation Equation assigns.

The Modified Adam & Eve case shows that we cannot use the Observation Equation to evaluate a conditional credence $P_\alpha(h|e)$ if e is consistent with more than one reference class. If a correct reference class definition can be found, then such a reference class definition must place any two observer-moments with the same total evidence in the same reference class.

Thirdly, Bostrom does not tell us how to divide observers into observer-moments. He only says that all observer-moments must have the same length in time, e.g. they must all be 1 hour long.[19] All the thought-experiments that we have considered so far that the Observation Equation could be successfully applied to (Modified Incubator, Adam and Eve, UN++ and

[19] Bostrom, p. 164.

Quantum Joe) did not require us to split observers up into observer-moments, but could be adequately represented by treating each observer as a single observer-moment. Moreover, although Bostrom says that all observer-moments must have the same length in time, this did not seem to be important when applying the Observation Equation to these thought-experiments. For example, the observers on the planet in the Modified Incubator could all have had lives that were not equal in length, but this would not have made a difference to their credences. Even in cases where an observer's total evidential state evolved over time, as for example in the Modified Incubator when the observers in the incubator exited their cells and observed their cell number, we did not have to represent the observers as consisting of two observer-moments, one in the cell and one outside of the cell; instead, in order to find their credence in heads while they were inside the cell, we evaluated $P_{\alpha}(heads|incubator)$, and in order to find their credence in heads once they exited the cell and observed that they were in, say, cell #1, we evaluated $P_{\alpha}(heads|cell\ \#1)$. This raises the question of whether observer-moments are at all necessary or if they can simply be replaced by observers. If they are necessary, then we must know under what conditions a single observer should be represented by more than one observer-moment in order to avoid imprecision in the theory.

Lastly, a source of imprecision for Bostrom's theory is that the credence function P_{α} is indexed to an observer-moment α, suggesting that different observer-moments can or should have different prior credence functions. I think that this is undesirable for two reasons. First, in trying to develop a theory of self-locating credences, we are trying to find norms of rationality. If we are trying to find norms of rationality to constrain credence functions, then these norms should constrain all observer-moments' credence functions. Therefore, unless there is good reason to think that different observer-moments should have different credence functions, and there does not appear to be any such good reason, a theory of self-locating credences should not allow for different observer-moments to have different prior credences in worlds under consideration, as this adds an unnecessary source of imprecision.

Now, I am not advocating the much stronger view that for *any* possible world *w*, all actual observers should have the same prior credence in *w*. That is far outside of the scope of what we are doing here. Rather, I am saying that in any possible world under consideration, all observer-moments under consideration should have the same prior credence function. Arntzenius and Dorr's Proportion, because it refers to a single prior credence function C, also implies that all observer-moments under consideration should have the same prior credences in propositions under consideration, so this is not a strong assumption. Hence, there is no reason why the Observation Equation should allow different observer-moments to have different prior credences.

Secondly, when we applied the Observation Equation to all of the above thought-experiments, we set the prior credences in worlds to be equal to the objective probability of those worlds, and the prior credence function was therefore the same for all observer-moments. In the Incubator and the Modified Incubator, for example, we set the prior credence in the heads-world to be 1/2, and in Quantum Joe we set the prior credence in the output of the quantum device being 0 to be 1/10. In Adam & Eve we did not specify the prior credence that conception will have occurred, but we obtained the result that Adam's and Eve's credence in conception is equal to the prior credence in conception, which is a desirable result just in case the prior credence in conception is equal to the objective probability of conception. Therefore, it seems that the Observation Equation should be supplemented with some version of the Principal Principle in order to place the credence function on firmer ground and avoid any potential problems stemming from the fact that the Observation Equation relates only conditional credences to prior credences.

In the next section, I will present my own theory of self-locating credences, which will be based on Bostrom's framework. I will show how my theory solves the problems and sources of imprecision in the Observation Equation that I have discussed above.

My theory

In this section, I will outline my own theory of self-locating credences. My theory consists of two principles, a modification of the Observation Equation and the Principal Principle.

Let C be a prior credence function, let *h* and *e* be propositions, let Ω_h and Ω_e be sets of observer-moments that *h* or *e* is true of respectively, and for any observer-moment σ let Ω_σ be σ's reference class, let w_σ be the world that σ is in and let $\Omega(w_\sigma)$ be the set of all observer-moments in w_σ. Then, the Observation Equation should be modified in the following way:

$$\mathrm{C}(h|e) = \frac{1}{\gamma} \sum_{\sigma \in \Omega_h \cap \Omega_e} \frac{\mathrm{C}(w_\sigma)}{|\Omega_\sigma \cap \Omega(w_\sigma)|}, \gamma = \sum_{\sigma \in \Omega_e} \frac{\mathrm{C}(w_\sigma)}{|\Omega_\sigma \cap \Omega(w_\sigma)|}$$

The modified Observation Equation is the same as the original Observation Equation, except that the credence function of observer-moment α, P_α, is replaced with a single prior credence function *C* that is the same for all observer-moments. Of course, this does not negate the fact that different observers have different posterior credences, as different observers have different evidence, including different beliefs about the objective probabilities of worlds. And, again, it must be stressed that this applies only to observer-moments under consideration in possible worlds under consideration. It is not intended as a general claim about actual people's prior credences in all possible worlds.

The Principal Principle is essentially the same as in Arntzenius and Dorr's theory. If C is a prior credence function and *w* is a possible world, then the Principal Principle states the following:

$$C(w|P \text{ is the objective chance function}) = P(w)$$

Now, I will outline my theory of reference classes. For now, we will suppose that every observer can be adequately represented as just one observer-moment, regardless of how long the observer in question lives. Later, I will explain when and how we should split up a single observer into more than one observer-moment. Hence, for now I will only use the term 'observer' for simplicity.

In order to understand reference classes, we must take note of the fact that the same observer can exist in more than one world. For example, suppose that I decide to toss a coin five seconds from now. Then, there are two worlds—a heads-world and a tails-world—corresponding to the possible outcomes of the coin toss. Crucially, however, the observer who tosses the coin in the heads-world and the observer who tosses the coin in the tails-world are the same observer. Importantly, the observer who tosses the coin in the heads-world and the observer who tosses the coin in the tails-world are the same observer in virtue of their causal history and not in virtue of any qualitative property. There is no intrinsic property, be it DNA or anything else, that is essential to the observer who tosses the coin and which would make it the case that the observer who tosses the coin in the heads-world is the same being as the observer who tosses the coin in the tails-world. I am making this stipulation for the following reason. There could be a possible world with a radically different universe from ours in which the extremely unlikely event happens of a being with my exact DNA spontaneously coming into existence, into completely different circumstances from my own actual circumstances. Or, purely by extreme chance, it is possible that someone with roughly my exact DNA could have been born during some other era in history. But because these possibilities are so extremely unlikely and completely causally disconnected from who I am in the actual world, we would not wish to say that these possible observers are the same being as me.

Instead, what is essential to me is the causal history of how I came into existence. After I was conceived, there are many ways in which my life could have gone completely differently, and there are possible worlds corresponding to the various ways in which my life could have gone (including, for example, worlds in which I died before I was born). What makes me me in those worlds is that the world up to my conception was sufficiently similar to the actual world, so that the causal process by which I was conceived occurred in more or less the same way. This is why, if I decide to toss a coin five seconds from now, it is I who is tossed the coin both in the heads-world and in the tails world. Indeed, it is not only I, an observer, who exist in both the heads-world and the tails-world. The coin that I toss is the same coin in the heads-world and the tails-world; all the objects on my desk and in my room exist in both worlds; if the coin toss occurs at a time *t*, the totality of the universe until time *t* and those parts of the universe after time *t* that are causally unaffected by the outcome of the coin toss, are the same in both possible worlds.

Preliminary reference class definition

Now, we are ready to state a preliminary reference class definition. It is only a preliminary definition because it is possible for two observer-moments with the same total evidential state to be in different reference classes according to this preliminary definition.

The condition for belonging to the same reference class is this: for any two observer-moments σ and τ, σ and τ belong to the same reference class if and only if for any world w under consideration, either σ and τ both exist in w or neither σ nor τ exists in w.

It is important to stress that this criterion only refers to possible worlds under consideration, rather than all possible worlds. For example, there is no other observer-moment who exists in all and only those possible worlds that I exist in. But if we are considering only possible worlds where, say, I and both of my siblings have existed, then my siblings and I are in the same reference class.

In order to see that this reference class definition is correct (under the assumption that every observer consists of a single observer-moment within any world), let us apply it to the previously discussed thought-experiments.

Incubator: There are 100 observers in the heads-world and 1 observer in the tails-world. The question is: does the observer who exists in the tails-world also exist in the heads-world? There is nothing in Incubator that would suggest that this is the case. The observer or observers are said to be created by the coin toss said. We might explicate this in the following way. There is a kind of causal process, called an incubation, that can be triggered in the incubator. Heads causes 100 incubations to be triggered, while tails causes only 1 incubation to be triggered. There is nothing that would suggest that the incubation process that is triggered in the event of tails is the same incubation as any of the incubations that occur in the heads-world. Therefore, we may conclude that the observer in the tails-world exists only in the tails-world while the observers in the heads-world all exist only in the heads-world. (Of course, one could fill in details in the Incubator to make it the case that the incubation in cell #1 is the same incubation in the heads-world and the tails-world; this would then put the observer in cell #1 into a separate reference class.)

Therefore, we have a reference class consisting of observers in the heads-world and a reference class consisting of observers in the tails-world. While they are inside their cell, no observer knows what reference class they belong to. This is not a problem when it comes to applying the (Modified) Observation Equation, because there is no observer that could be in more than one reference class *within a single world*. The prior credence in heads is 1/2, equal to the objective probability of heads. It is easy to verify that this reference class partition will give us the desired results: C(*heads*) = 1/2 and C(*heads*|*cell* #1) = 1/101.

Modified Incubator: There are two groups of observers—the planet-observers and the incubator-observers. We have seen from the original Incubator case that the observer in the incubator in the tails-world does not exist in the heads-world. The planet-observers exist in both the heads-world and the tails-world, as their existence is causally independent of the outcome of the coin toss. Therefore, we have three reference classes: the planet-observers, the incubator-observers in the heads-world and the incubator-observer in the tails-world. The prior credence in heads is again 1/2, by the Principal Principle. It is easy to verify that this reference class partition will give us the desired results: C(*heads*|*planet*) = 1/2, C(*heads*|*incubator*) = 1/2, C(*heads*|*cell* #1) = 1/101.

The remaining thought-experiments Adam and Eve, UN++ and Quantum Joe follow the same pattern. In each of these thought-experiments, there is a reference class consisting of observers who exist in all possible worlds—Adam and Eve, the Earth inhabitants and Joe, respectively. Adam's and Eve's credence in conception will be equal to the prior credence in conception, the Earth inhabitants' credence that any particular gamma ray burst will occur will be 9/10, which is the prior credence in and objective probability of any given gamma ray burst occurring, and Joe's credence in the quantum device giving the output 0 is 1/10, again equal to the prior credence in 0 and the objective chance of 0. In general, if we have a reference class Ω_0 of observers that exist in every world, C is the credence function and w is any world, then $\mathrm{C}(w|\Omega_0) = \mathrm{C}(w)$; that is, the credence in any world conditional on being in that reference class will equal the prior credence in that world. This is easy to show:

$$\mathrm{C}(w|\Omega_0) = \frac{\sum_{\sigma\in\Omega(w)\cap\Omega_0} \frac{\mathrm{C}(w)}{|\Omega(w)\cap\Omega_0|}}{\sum_{\sigma\in\Omega_0} \frac{\mathrm{C}(w_\sigma)}{|\Omega(w_\sigma)\cap\Omega_0|}} = \mathrm{C}(w)$$

As we have seen, our preliminary reference class definition says that which reference class an observer-moment belongs to depends on which other observers exist in all and only the same possible worlds as them. Therefore, in order to know his or her reference class, an observer-moment has to know which possible worlds he or she exists in. Secondly, which possible worlds an observer-moment exists in depends on facts about their causal history. Facts about an observer's causal history are not at all constrained by that observer-moment's total evidence.

For example, suppose that we have an observer called Bob and that a coin is tossed before Bob is born. If the coin lands heads, nothing is done to Bob and he is allowed to be born and live a normal life without any intervention from anyone else. If the coin lands tails, Bob is taken immediately after birth and is placed into a vat that feeds experiences directly to his brain. These experiences are as of a world that is radically different from the actual world; they convince him that he is an intelligent lizard-like being living on a planet completely different from our own in a galaxy completely different from our own. Clearly, Bob's experiences give him a completely different set of beliefs about his causal history, and therefore also about his identity and reference class membership as we have defined it thus far, in the heads-world than they do in the tails-world. There is nothing in our experiences that can tell us facts about our causal history with certainty, because there is always the possibility that our experiences do not match the underlying reality of our causal history—and as we have defined it thus far, reference class membership depends only on the underlying facts about our causal history.

But we have already seen that the Observation Equation gives strange results if we try to conditionalise on evidence that is consistent with being in more than one reference class within the same world. Therefore, the preliminary reference class definition is liable to counter-examples where an observer's total evidential state during some period in time is consistent with more than one reference class within the same world. We already saw an example of this in the Modified Adam and Eve thought-experiment. For every observer, there is a period of time (namely their 20th birthday) when they do not know whether they are Adam or Eve or one of their descendants, since their eyes are closed and their memory is erased during this time. If we place all temporal parts of Adam and Eve in one reference class and all temporal parts of their descendants in another reference class and try to evaluate the Observation Equation conditional on T (where T is a tautology and Ω_T therefore includes all observers), we get a nonsensical result.

Before I outline my solution to this problem, we should let us consider two more thought-experiments, *Incubator** and *Incubator***, which exemplify the same problem and will support my solution to it.

> **Incubator*:** An observer is created in cell #1 and is initially asleep. A fair coin is then tossed. If it lands heads, 99 more observers are created, one in each of cells #2–100, and are initially asleep. If it lands tails, no new observers are created. Any existing observers are then awakened. After one hour, they are released from their cells and observe what cell they were in. All observers know the setup. What should an observer's credence in heads be while they are inside the cell? What should their credence in heads be if they observe that they were in cell #1 when they leave the cell?
>
> **Incubator**:** An observer is created in cell #1. After one hour, they are released from the cell and observe that they were in cell #1. One hour later, a fair coin is tossed. If it lands heads, 99 more observers are created, one in each of cells #2–100. One hour after that, they are released from their cells and observer what cell they were in. All observers know the setup. What should an observer's credence in heads be while they are inside the cell? What should their credence in heads be if they observe that they were in cell #1 when they leave the cell?

Let us start with Incubator**. What should C(*heads*|*cell* #1) be in the Incubator** case? I think that it should be 1/2. The observer who finds out that they were in cell #1 knows that a fair coin toss will occur one hour in the future. Therefore, I think that it would be wrong to claim that their credence in heads should be anything other than 1/2. This is the credence that my preliminary reference class definition dictates. The observer in cell #1 is the same in the heads-world and the tails-world. Therefore, since this observer exists in all of the possible worlds under consideration, their credence in heads conditional on knowing what reference class they are in should be equal to the prior credence in heads, which should be equal to the objective probability of heads, which is 1/2. Arntzenius and Dorr would have to say that C(*heads*|*cell* #1) = 1/101, because in light of Proportion and the Principal Principle the Incubator** thought-experiment does not have any relevant differences from the original Incubator thought-experiment. Hence, Arntzenius and Dorr's treatment of Incubator** parallels their treatment of Adam and Eve, UN++ and Quantum Joe, where Adam and Eve, the Earth inhabitants and Joe have credences in outcomes of chance events that are different from the objective chances of these outcomes, despite the fact that they exist independently of the outcomes.

Now, let us consider Incubator*. What should C(*heads*|*cell* #1) be in the Incubator* case? I think that it should also be 1/2. The only difference between Incubator* and Incubator** is that when the observer in cell #1 leaves his cell, he knows that the coin toss has already occurred. Since the causal process by which the observer in cell #1 came into existence is independent of the outcome of the coin toss, when the observer in cell #1 exits the cell and observes that he was in cell #1, he also knows that he was necessarily in cell #1; that whatever the outcome of the coin toss, he would have been in cell #1. Therefore, after he observes that he was in cell #1, his posterior credence in heads must be 1/2. This is what my preliminary reference class definition predicts, since the observer in cell #1 exists both in the heads-world and in the tails-world while the other observers all exist only in the heads-world. If we accept that C(*heads*|*cell* #1) = 1/2 in Incubator**, then claiming that the value of C(*heads*|*cell* #1) is different in the case of Incubator* would seem to commit us to a strange view according to which the time at which an event occurs is relevant to an observer's credence in that event even if the observer has not in any way been causally influenced by that event.

Therefore, in both Incubator* and Incubator**, C(*heads*|*cell* #1) = 1/2. Obviously, C(*heads*|*cell* #*n*) = 1 for all *n* such that 1 < *n* < 101 in both Incubator* and Incubator**. The difficult question is the following: if '*incubator*' denotes the proposition that I am in the incubator, what should the value of C(*heads*|*incubator*) be? That is, what should an observer's credence in heads be while they are still inside the cell?

I think that an observer's credence in heads while they are inside the cell should be 1/2, in both Incubator* and Incubator**, for the following reason. Every observer will have had the same experiences and the same total evidence throughout the first hour during which they are awake—namely, the experience of being in the incubator and the evidence that they are in the incubator. Therefore, the experience of being inside a cell in the incubator cannot contain any evidence that would allow the observer to shift their credence in heads away from the objective probability of heads, which is 1/2. The only evidence that the experience of being inside the incubator contains is that at least one observer exists inside the incubator. But given either heads or tails, it is certain that at least one observer exists inside the incubator. Therefore, the observer inside the incubator has no evidence that would allow them to shift their credence in heads away from 1/2.

Arntzenius and Dorr would agree that an observer inside the incubator should have credence of 1/2 in heads both in the Incubator* case and in the Incubator** case—but this is only because they would also say that an observer's credence in heads conditional on being in cell #1 should be 1/101. So, one possible objection to my view that C(*heads*|*incubator*) = 1/2 can be set out in the following way:

$$\begin{aligned} \mathrm{C}(heads|incubator) \\ &= \mathrm{C}(heads|cell\ \#1)\mathrm{C}(cell\ \#1) \\ &+ \mathrm{C}(heads|cell\ \#2\ or\ \#3\ or\ \ldots or\ \#100)\mathrm{C}(cell\ \#2\ or\ \#3\ or\ \ldots or\ \#100) \end{aligned}$$

Note that C(*heads*|*cell* #2 *or* #3 *or* … *or* #100) C(*cell* #2 *or* #3 *or* … *or* #100) = 1 – C(*cell* #1). If C(*heads*|*incubator*) = C(*heads*|*cell* #1) = 1/2, then we obtain the following:

$$\frac{1}{2} = \frac{\mathrm{C}(cell\ \#1)}{2} + 1 - \mathrm{C}(cell\ \#1)$$

$$1 = 2 - \mathrm{C}(cell\ \#1)$$

$$\mathrm{C}(cell\ \#1) = 1$$

This would imply that an observer in the incubator must be certain that they are in cell #1, which is obviously not the case. Hence, the observer in cell #1 does not use Bayesian updating to update his credences when they exit the cell and discover that they were in cell #1.

However, the above equation must be rejected. It is indeed the case that in the Incubator* and the Incubator** thought-experiments, the reasoning of the observer in cell #1 cannot be represented by Bayesian conditionalisation. However, this is not only not a problem, but is in fact to be expected, for the following reason. When an observer is released from the cell and learns that they were in cell #1, they do not only learn that they were in cell #1. They learn something more than that: they learn that they were *necessarily* in cell #1; they learn that they had to exist, that they had to be in cell #1 and that they would have observed that they were in cell #1, all regardless of the outcome of the coin toss. Note that the same cannot be said of the observers in cells #2–100. When they are released from their cells and when they learn that they were not in cell #1 but in some other cell, they also learn that they were only contingently in whichever cell they were in, and that, had the outcome of the coin toss been tails instead of heads, they would not have existed in any cell.

What specifically goes wrong with the above equation for C(*heads*|*incubator*), and how can we explain our reasoning in the Incubator* and Incubator** cases? What goes wrong is that there is no single prior credence function that obeys the probability axioms and which represents an observer's credences both before and after they are released from the cell. Let C_0 denote the credence function of any observer before they are released from their cell. Then, C_0(*heads*|*incubator*) = 1/2 as we have said, but also C_0(*cell* #1|*heads*) = 1/100. However, once the observer exits the cell and finds that they were in cell #1, their credence function changes. If C_1 denotes the posterior credence function of the observer in cell #1, then C_1(*cell* #1) =1 and C_1(*heads*) = 1/2.

Instead, we must treat the Incubator* and the Incubator** in the following way. Every observer consists of two observer-moments: an early observer-moment, which is inside a cell in the incubator, and a late observer-moment, which has exited their cell and observed their cell number. The late observer-moments are grouped in two reference classes: Ω_1, which contains the observer in cell #1 in the heads-world w_1 and the tails-world w_2, and Ω_2, which contains the other observers. The early observer-moments all belong to a single reference class, Ω_0. If C denotes the prior credence function, $C(w_1) = C(w_2) = 1/2$, since these are the objective probabilities of heads and tails. Previously, we used '*incubator*' to denote a proposition that included all observer-moments, because all observers were represented by only one observer-moment and every observer was inside a cell in the incubator for a certain temporal part of its existence. Now, we will use '*incubator*' to mean that the observer is currently in the incubator. Hence, '*incubator*' will refer only to the early observer-moments and will be synonymous with Ω_0. Likewise, we will use '*cell* #1' to denote the proposition that the observer is observing that they were in cell #1 and not that they are actually in cell #1. Hence, '*incubator*' will be true of all and only the observer-moments in reference class Ω_1, and will not contain any observer-moments from Ω_0. Then, we get the desired credence in heads of 1/2 for observers while they are in the incubator and the observer in cell #1 after he exits the cell:

$$\mathrm{C}(heads|incubator) = \frac{\sum_{\sigma\in\Omega(w_1)\cap\Omega_0} \frac{\mathrm{C}(w_1)}{|\Omega(w_1)\cap\Omega_0|}}{\sum_{\sigma\in\Omega_0} \frac{\mathrm{C}(w_\sigma)}{|\Omega(w_\sigma)\cap\Omega_0|}}$$

$$= \frac{\sum_{\sigma\in\Omega(w_1)\cap\Omega_0} \frac{\mathrm{C}(w_1)}{|\Omega(w_1)\cap\Omega_0|}}{\sum_{\sigma\in\Omega(w_1)\cap\Omega_0} \frac{\mathrm{C}(w_1)}{|\Omega(w_1)\cap\Omega_0|} + \sum_{\sigma\in\Omega(w_2)\cap\Omega_0} \frac{\mathrm{C}(w_2)}{|\Omega(w_2)\cap\Omega_0|}} = \frac{\mathrm{C}(w_1)}{\mathrm{C}(w_1)+\mathrm{C}(w_2)}$$

$$= \frac{1}{2}$$

$$\mathrm{C}(heads|cell\ \#1) = \frac{\sum_{\sigma\in\Omega(w_1)\cap\Omega_1} \frac{\mathrm{C}(w_1)}{|\Omega(w_1)\cap\Omega_1|}}{\sum_{\sigma\in\Omega_1} \frac{\mathrm{C}(w_\sigma)}{|\Omega(w_\sigma)\cap\Omega_1|}}$$

$$= \frac{\sum_{\sigma\in\Omega(w_1)\cap\Omega_1} \frac{\mathrm{C}(w_1)}{|\Omega(w_1)\cap\Omega_1|}}{\sum_{\sigma\in\Omega(w_1)\cap\Omega_0} \frac{\mathrm{C}(w_1)}{|\Omega(w_1)\cap\Omega_1|} + \sum_{\sigma\in\Omega(w_2)\cap\Omega_0} \frac{\mathrm{C}(w_2)}{|\Omega(w_2)\cap\Omega_1|}} = \frac{\mathrm{C}(w_1)}{\mathrm{C}(w_1)+\mathrm{C}(w_2)}$$

$$= \frac{1}{2}$$

Now, let us consider Modified Adam & Eve. Recall that Modified Adam & Eve is the version of Adam & Eve in which observers have their eyes closed and memory erased on their 20th birthday. We can treat Modified Adam & Eve similarly to Incubator* and Incubator**. First, we divide every observer into two observer-moments: one observer-moment is the observer's 20th birthday, and the other observer-moment is the time before and after the observer's 20th birthday. (This shows that observer-moments can be discontinuous in time.) The observer-moments that comprise an observer's 20th birthday will comprise one reference class, Ω_0. The other observer-moment of Adam and of Eve in the conception-world w_1 and the no-conception-world w_2 will comprise a second reference class Ω_1, and the other observer-moments of Adam and Eve's descendants will comprise a third reference class Ω_2. It is easy to verify that given this reference class partition, $\mathrm{C}(w_1|\Omega_1) = \mathrm{C}(w_1)$ and $\mathrm{C}(w_1|\Omega_2) = \mathrm{C}(w_1)$. In words, Adam's and Eve's credence in conception will be equal to the objective probability of conception, and the closed-eyed observer-moments' credence in conception will also be equal to the objective probability of conception. I think that this is the desirable result.

This result for Modified Adam & Eve might seem undesirable. Specifically, it might seem that the closed-eyed observer-moments' credence in conception should be much higher than the objective probability of conception. However, I think that this intuition is mistaken. During their 20th birthday, every observer, including Adam and Eve, is in the same total evidential state. All one knows in this state is that someone exists, which is not evidence for or against conception and hence cannot shift an observer's credence in conception away from the objective probability of it. When we imagine the Modified Adam & Eve scenario, our intuitions may lead us astray because when we imagine what it would be like to be a closed-eyed observer-moment, we imagine ourselves with our eyes closed and memories intact, even if we try otherwise. Moreover, our intuition is led astray because the actual world is a 'conception-world' (in that it will have contained hundreds of billions of observers) and because we may imagine the *prior* credence of conception to be very high. On this last point, it is important to keep in mind that even if the prior credence in conception is very high, what we are talking about is whether a closed-eyed observer-moment's credence in conception is any higher than the prior credence in conception, or equal to it.

A further reason why the closed-eyed observer-moments in Modified Adam & Eve have to have credence in conception equal to the objective probability of conception is because we want to avoid the view that worlds with a greater number of observers should be weighed more heavily in self-locating reasoning. Arntzenius and Dorr as well as Bostrom reject any such view, arguing that our prior probability in a world cannot be weighed more if it has more observers, even if the number of observers is greater by a factor of a trillion. Bostrom argues for this by describing the following scenario:

> **The Presumptuous Philosopher:** ‘It is the year 2100 and physicists have narrowed down the search for a theory of everything to only two remaining plausible candidate theories, T1 and T2… According to T1 the world is very, very big but finite, and there are a total of a trillion trillion subjects in the cosmos. According to T2, the world is very, very, very big but finite, and there are a trillion trillion trillion subjects. The super-duper symmetry considerations are indifferent between these two theories. Physicists are preparing a simple experiment that will falsify one of the theories. Enter the presumptuous philosopher: “Hey guys, it is completely unnecessary for you to do the experiment, because I can already show to you that T2 is about a trillion times more likely to be true than T1!”.’[20]

This scenario shows that if we held the view that we should weigh our credences in worlds by the number of observers they contain, then the presumptuous philosopher’s conclusion would be valid. Yet it is hard to believe that this could be the case, because it would mean that we can draw strong empirical conclusions in physics a priori. The scenario as described assumes that the physicists’ prior credence in each of T1 and T2 is approximately 1/2.

To make the scenario even more vivid, we could change it in the following way. We could say that the physicists have already developed a theory of everything, T0. T0 says that roughly 13.4 billion years ago, a chance event occurred that would determine the nature of the universe from that point onwards. The chance event had two possible outcomes, T1 and T2, each of which had an objective probability of 1/2. The outcome T1 would have caused the universe to be very, very big but finite and to contain a trillion trillion subjects, while the outcome T2 would have caused the universe to be very, very, very big but finite and to contain a trillion trillion trillion subjects. In that case, the presumptuous philosopher’s claim would be even less tenable, as it would mean that the fact that T2 predicts a universe with more

[20] Bostrom, p. 124.

subjects would override not only symmetry considerations, but the objective probabilities of the two hypotheses. The intuition that the closed-eyed observer-moments' credence in conception should be different from the objective probability of conception is exactly analogous to and just as mistaken as the presumptuous philosopher's intuition about what our credences in T1 and T2 should be.

In this section, I have shown how we can split observers into observer-moments in cases where there is a period of time during which an observer's total evidence is not sufficient for them to know which reference class they are in according to the preliminary reference class definition that I proposed at the beginning of this section. I have not yet given a principled set of rules for how this should be done in general, however. This is what I will do in the next section.

General reference class definition

In this section, I will begin by specifying exactly how references classes should be delineated, as well as how observers should be divided into observer-moments. Then, I will explain why I think this reference class definition is correct.

I have already proposed a preliminary reference class definition: for any two observer-moments σ and τ, σ and τ are in the same reference class if and only if for any world w, either σ and τ both exist in w or neither σ nor τ exist in w; equivalently, σ and τ are in the same reference class if and only if there is no world w such that exactly one of σ and τ exists in w.

Why is this preliminary reference class definition not generally adequate? It is not generally adequate only because of the possibility that some observers, at some times, may not know what preliminary reference class they are in. To put the same point more precisely, the preliminary reference class definition is not adequate because of the following possibility: there is a possible world w, a total evidential state e, and observers σ and τ such that σ and τ

are not in the same reference class and such that they are both in state *e* in world *w* at some point in time. Intuitively, what this means is that there is some total evidential state that is consistent with being in more than one reference class in the same world. We saw examples of this in Incubator*, Incubator** and Modified Adam & Eve. In Incubator* and Incubator**, the total evidential state that observers are in before they are released from their cell is consistent both with being the observer in cell #1 and with being any of the observers in cells #2–100; and according to the preliminary reference class definition, the observer who is in cell #1 is in a different reference class from the observers who are in cells #2–100. In Modified Adam & Eve, the total evidential state that observers are in during their 20th birthday is consistent with being any of the observers, and the preliminary reference class definition puts Adam and Eve in one reference class, and their descendants in another.

In order to find any given observer-moment's reference class, we follow a two-stage procedure. In the first stage, we sort all observer-moments into preliminary reference classes according to the preliminary reference class definition. To be more precise, let σ and φ be arbitrary observer-moments possibly in different worlds, let Σ denote the observer that σ is a part of and let Φ denote the observer that φ is part of. Then, φ belongs to σ's preliminary reference class O_σ if and only if for any world *w*, either Σ and Φ both exist in *w* or neither Σ nor Φ exists in *w*.

In the second stage, we do the following. Let σ be any observer-moment, let O_σ denote σ's preliminary reference class (as defined in the first stage), and let *e* denote a proposition that describes σ's total evidential state. Ω_e denotes the set of observer-moments that *e* is true of—that is, Ω_e is the set of all observer-moments φ such that σ and φ are in the same total evidential state. Now, if Ω_e is a subset of O_σ—that is, if all observer-moments with σ's total evidence are part of σ's preliminary reference class—then σ's reference class Ω_σ is the same as its preliminary reference class, O_σ. On the other hand, if Ω_e is not a subset of O_σ—that is, if there are observer-moments with σ's total evidence that are outside of σ's preliminary reference class—then for any observer-moment φ that is in the total evidential state e, φ's preliminary reference class O_φ must be a subset of σ's reference class Ω_σ.

In other words, if an observer-moment's total evidence narrows it down to a single preliminary reference class, then that observer-moment's reference class is the same as its preliminary reference class. On the other hand, if an observer-moment's total evidential state is consistent with more than one preliminary reference class, then that observer-moment's reference class is equal to the union of all preliminary reference classes that its total evidential state is consistent with.

Now, let us discuss some of the consequences of this reference class definition. The first thing we should note is that while the set of preliminary reference classes is a partition over the set of all observer-moments, the set of reference classes is not a partition over the set of all observer-moments. The fact that the set of preliminary reference classes is a partition over the set of all observer-moments means the following: (1) every observer-moment σ belongs to its own preliminary reference class O_σ, and (2) for any observer-moments σ and τ, if τ belongs to σ's preliminary reference class O_σ, then τ's preliminary reference class O_τ and O_σ are the same set, and σ belongs in O_τ. In other words, every observer-moment belongs to exactly one preliminary reference class.

On the other hand, the set of reference classes is not a partition over the set of all observer-moments. This is because although condition (1) from the previous paragraph holds–every observer-moment σ belongs to its own reference class Ω_σ—condition (2) does not hold, because it is possible for some observer-moment τ to belong to another observer-moment σ's reference class Ω_σ without σ belonging to τ's reference class Ω_τ. For example, suppose that σ and τ are observer-moments who have total evidential state e and f respectively, such that both observer-moments' preliminary reference classes O_σ and O_τ contain observer-moments that are in the evidential state *f* but τ's preliminary reference class O_τ does not contain any observer-moments that are in the evidential state *e*. Then, σ is contained in τ's reference class Ω_τ but τ is not contained in σ's reference class Ω_σ. However, reference classes are not totally disorganised; they do have a certain structure. Specifically, for any reference class Ω_τ and for any pair of observer-moments σ_1 and σ_2 that belong to the same preliminary reference class O_σ, it will either be the case that both σ_1 and σ_2 are contained in Ω_τ or that neither σ_1 nor σ_2 is contained in Ω_τ. In other words, given any preliminary reference class O_σ and reference class Ω_τ, it will either be the case that O_σ is a subset of Ω_τ or it will be the case that O_σ and Ω_τ are disjoint. Hence, reference classes are unions of preliminary reference classes.

From the previous paragraph, a visual picture of self-locating reasoning emerges that clarifies how self-locating reasoning works. If we ignore reference classes, as Arntzenius and Dorr do, and effectively assume that there is only one reference class, then we have a credence function that assigns to any uninhabited world a credence of 0, assigns to the proposition that one is any observer-moment in a given inhabited world a credence proportional to the objective chance of that world, and finally assigns the same credence to every observer within a given world. As an observer-moment's evidential state evolves, various observers in various worlds that are inconsistent with the observer-moment's total evidence get ruled out (and some might get ruled in, in cases of forgetting), and the credence in each remaining observer in each world gets a proportional increase (or decrease). The credence function therefore always changes in accordance with Bayesian conditionalisation (except in cases of forgetting).

The picture I have described differs in just one way. Whenever the observer's total evidential state evolves in such a way that some preliminary reference class is ruled out (or, for example due to forgetting, an observer is ruled in whose entire preliminary reference class was previously ruled out), credences in individual observer-moments do not get proportionally increased or decreased. Instead, let *e* be the total evidential state that causes a preliminary reference class to be ruled out for the first time. Then, the credence function first assigns the credence 0 to all observer-moments in any preliminary reference class that is inconsistent with *e*. Then, the function assigns to each remaining world (i.e. world that has preliminary reference classes consistent with *e*) a credence proportional to the prior credence in that world. Then, in any given world, each remaining observer-moment (i.e. each observer-moment in a preliminary reference class consistent with *e*) is assigned the same credence. Finally, this assignment of credences is conditionalised on *e*. When and only when this happens, the observer violates Bayesian conditionalisation. However, I think that this is a desirable result, because the observer has not merely acquired new (or lost old) evidence about who they might be in the world, but also about what facts their existence depends on, and therefore what kinds of evidence and what kinds of experiences they could have had, could not have had, had to have and had not to have.

Before continuing, let us briefly consider again the Incubator*, Incubator** and Modified Adam & Eve cases. In each of these cases, we split every observer into an 'early' and a 'late' observer-moment. In the Incubator* and the Incubator**, the early observer-moments were the ones inside their cells while the late observer-moments were the ones who had come out of their cells. In Modified Adam & Eve, the early observer-moments were the temporal parts of all observers during their 20th birthday who had their eyes closed and their memories erased, while the late observer-moments were the remaining parts of all observers, both before and after their 20th birthday (showing that observer-moments do not have to be continuous in time), who had their eyes open and knew who they were. When we previously applied the Observation Equation to these thought-experiments, we placed all of the early observer-moments in one reference class that did not include any of the late observer-moments. By contrast, the reference class definition that I have now given dictates that the early observer-moments' reference class should also include all of the late observer-moments and that each late observer-moment's reference class should also include some of the early observer-moments, for the following reason: for any given observer, their early observer-moment and their late observer-moment either both exist or both do not exist in every world; therefore, they have to be in the same preliminary reference class; therefore, the early observer-moments' reference class has to include all observer-moments and each late observer-moment's reference class has to include some early observer-moments.

Two things should be said about this. First, and most importantly, which type of reference class partition we use to represent cases like these will not affect observer-moments' credences. This is easy to verify, so I will not do it here. I used the first kind of reference class partition for simplicity, as my goal then was only to show that cases where an observer does not know their preliminary reference class do not have to pose a problem for Bostrom's theory. Secondly, the reason why it does not matter how we partition reference classes in any of these cases is because there is no possible observer-moment that does not know whether they are an early observer-moment or a late observer-moment. The early observer-moments are not completely without evidence—they know that they are early. Therefore, since no observer-moment in any of these cases can be in a state of confusion about whether they are early or late, as their total evidential state always tells them this, the 'coarser' and the 'finer' reference class partitions work equally well.

Explanation of the general reference class definition

In the section before the last, I proposed a preliminary reference class definition according to which two observers belong to the same reference class if and only if there is no world that contains only one of the two observers. The preliminary reference class definition has to be adjusted in order to account for cases in which an observer's total evidence is consistent with membership in more than one preliminary reference class. Nonetheless, the preliminary reference class definition is adequate when every observer's total evidence at every time is consistent with only one preliminary reference class. In this section, I will explain why this is the case. In doing so, it will become clear that the reference class definition is very powerful, and can explain not only the very simple thought-experiments of the kind that we have considered in this text, but any case that involves self-locating reasoning in general.

All of the thought-experiments that we have considered in this text have three common elements. First, they all have a chance event or a series of chance events. In all versions of the Incubator, this is the coin toss; in all versions of Adam & Eve, the possible conception. In UN++, there is a series of chance events, the chance events being possible gamma ray bursts. In Quantum Joe, the chance event is a quantum measurement. Secondly, some of the thought-experiments have a class of independent observers. These are observers whose existence is independent of the outcome of the chance event in question. The standard Incubator thought-experiment does not have a class of independent observers, but Modified Incubator does, these being the observers on the planet; Incubator* and Incubator** have a class of independent observers as well, this being the observer in cell #1. The independent observers in Adam & Eve are Adam and Eve, the independent observers for any given gamma ray burst in UN++ are the observers that exist before that gamma ray burst, and the independent observer in Quantum Joe is Joe. Lastly, all of the thought-experiments have a class of dependent observers. These are observers who exist only if the chance event has a particular outcome. In the original Incubator and Modified Incubator, these are the observers in the incubator, while in Incubator* and Incubator** these are the observers in cells #2–100; in Adam & Eve, these are the descendants of Adam and Eve; in UN++, these are the inhabitants of space colonies created after a particular gamma ray burst; and in Quantum Joe these are the clones of Joe.

Now, each outcome of a chance event corresponds to a possible world. Since the independent observers exist regardless of the outcome, they exist in all of the possible worlds. Any dependent observer exists in only one of the possible worlds, because their existence is dependent on a particular outcome of the chance event (except in the UN++, where the outcome of any gamma ray burst except the last is represented by more than one world). Hence, the preliminary reference class definition that I have proposed separates the independent observers from the dependent observers in any scenario that follows the pattern established by Incubator, Adam & Eve and others.

However, we do not want our reference class definition to only apply to one narrow, contrived sort of case—we want it to be generally applicable. For example, what if there are many chance events? Let us start with a simple case. Suppose that there are two chance events. The first chance event has two possible outcomes, *a* and *b*. Depending on the outcome of the first chance event, the second chance event will have either *m* or *n* possible outcomes. Then, we have $m + n$ possible worlds. Worlds $w_1, \dots, w_m$ are those in which the outcome of the first chance event is *a* (so that w_i is the world in which the outcome of the first chance event is *a* and the outcome of the second chance event is *i*), while worlds $w_{m+1}, \dots, w_{m+n}$ are those in which the outcome of the first chance event is *b* (so that w_{m+i} is the world in which the outcome of the first chance event is *b* and the outcome of the second chance event is *i*).

Now, suppose we have three groups of observers. One group consists of observers whose existence is causally independent of the outcome of the first chance event. Since the second chance event itself depends on the outcome of the first chance event, the existence of the observers in this first group is also causally independent on the outcome of the second chance event. A second group of observers consists of observers whose existence is dependent on the outcome of the first chance event, but independent of the outcome of the second chance event. Observers in this group come into existence if the outcome of the first chance event is a, or if the outcome of the first chance event is b. However, this is all that is required for them to exist, and the outcome of the second chance event does not play any part in determining whether the observers in the second group exist. Finally, there is a third group of observers. This group consists of observers whose existence depends on the outcome of the second chance event. Since the second chance event itself depends on the outcome of the first chance event, the observers in this third group are dependent both on the first chance event and on the second chance event. In other words, in order for any given observer in the third group to exist, both chance events have to have the right outcome. For example, an observer in the third group who exists in the world w_{m+3} exists only in that world; and in order for them to exist, it is necessary for the outcome of the first chance event to be b and for the outcome of the second chance event to be 3.

Clearly, the first group of observers all exist in all of the worlds $w_1, \dots, w_m, w_{m+1}, \dots, w_{m+n}$. Observers from the second group all exist either in all of $w_1, \dots, w_m$, or in all of $w_{m+1}, \dots, w_{m+n}$, but not both. Lastly, observers from the third group all exist in only one of the worlds $w_1, \dots, w_m, w_{m+1}, \dots, w_{m+n}$. Hence, just by looking at what worlds observers exist in, we can determine which chance events they are independent of or dependent on, and on which outcomes they are dependent if they are dependent. Intuitively, following the pattern set by thought-experiments like Incubator and Adam & Eve, we would assign the first group of observers to one reference class, the second group of observers to a second reference class and the third group of observers to a third reference class.

This is effectively what my preliminary reference class definition does. It would place all the observers in the first group into one preliminary reference class. It would place all the observers in the second group that are dependent on the outcome a into a second preliminary reference class, and would place those dependent on b into a third preliminary reference class.

Finally, it would place all the observers in the third reference class into $m + n$ different preliminary reference classes.

This reference class partition is effectively the same as simply sorting the observers into just three reference classes because reference class boundaries matter only within worlds. This can be seen from the fact that in the Observation Equation, an observer-moment σ's reference class Ω_σ only ever appears as part of the intersection with the set of observers in the world in which σ is located, i.e. as part of $\Omega_\sigma \cap \Omega(w_\sigma)$.

In this example, we had two chance events, with three groups of observers defined in terms of their independence of or dependence on the outcomes of these chance events. Clearly, however, this example generalises to arbitrarily many chance events with any number of outcomes. The general picture that comes out of this is that of a branching world, where branches represent outcomes of chance events that result in further chance events and in the creation of observers. Then, reference classes can be defined equally well in terms of what outcomes they are dependent on and what chance events they are independent of. An observer will be independent of any chance event that they are causally upstream from (such as possible future coin tosses) and will exist in all possible worlds that represent the outcomes of chance events that they are independent of. So, for example, if I am independent of a possible future coin toss (which itself may depend on the outcome of another chance event that I am independent of), then I exist in all of the worlds that represent the outcomes of that coin toss. On the other hand, my existence depends on a specific outcome of some past event, so I exist only in those worlds in which that outcome of that chance event occurs. In turn, the chance event that I am immediately dependent on as an observer may itself be dependent on a particular outcome of another chance event that came before. The observers who are in my preliminary reference class are those who are dependent on all the exact same outcomes as I am, and independent of all the exact same chance events as I am. Hence, the observers who are in my preliminary reference class exist in all and only those worlds in which I exist.

I think that this picture of a branching world is generally applicable to self-locating reasoning. Although I may have left some possibilities out (such as a chance event occurring in more than one but not all outcomes of a previous chance event, or one chance event in some way interacting with another chance event), I think cases like this do not threaten my preliminary reference class definition with convincing counter-examples.

A possible objection to this is that not all hypotheses that self-locating evidence is relevant to are strictly speaking chance events. For example, we may use self-locating evidence in reasoning about which of two physical hypotheses about the universe is more likely to be true. However, even when we are not literally talking about chance events, thinking in terms of a branching world with chance events is still a useful heuristic. For example, suppose that we are trying to decide the impact of our evidence on two physical theories about the universe, T_1 and T_2, and that we think that the objective chance of each of these two theories is 1/2. We can treat T_1 and T_2 as outcomes of a chance event that all observers in the universe are dependent on, despite the fact that there may not be a literal chance event of this kind.

Sleeping Beauty

In this section, I will discuss the Sleeping Beauty problem. In all of the thought-experiments that I have discussed so far, the only purpose of dividing observers into observer-moments was in order to be able to represent a change in an observer's reference class over time. However, situations in which some observers do not know what the present time is also require us to divide observers into observer-moments. The Sleeping Beauty problem is an example of such a case.

The Sleeping Beauty problem first appeared in 'Self-locating belief and the Sleeping Beauty problem' by Adam Elga. It can be summarised as follows.

> **Sleeping Beauty:** Sleeping Beauty is put to sleep on Sunday night, and a fair coin is then tossed. Whether the coin lands heads or tails, she is awakened, told that it is Monday after some time passes and then put back to sleep and has the memory of the Monday waking wiped. Additionally, if the coin lands tails, she is also awakened on Tuesday and then put back to sleep. Finally, she is awoken on Wednesday and the experiment ends. Sleeping Beauty is told all this will happen before she is put to sleep on Sunday. What should Sleeping Beauty's credences be on Sunday, when she is first woken up on Monday and after she is told that it is Monday?[21]

Let us now look at how Sleeping Beauty should be represented. There are two possible worlds under consideration: the heads-world and the tails-world. The heads-world contains three observer-moments, which represent Sleeping Beauty on Sunday, Sleeping Beauty when she is first awoken on Monday, and Sleeping Beauty after she is told that it is Monday. The tails-world contains four observer-moments, representing Sleeping Beauty when she is awoken on Tuesday in addition to the ones mentioned in the heads-world. We may assume that all of these observer-moments last the same amount of time.

The next question is the following: should we place all of these observer-moments into one reference class, or should we partition them in some way. According to the preliminary reference class definition and the general reference class definition given above, it might seem that all of the observer-moments should be placed in the same reference class, because there is only one observer, Sleeping Beauty, and she knows the setup and therefore knows which observer she is. However, if we do this, we obtain obviously incorrect credences. For example,

[21] Elga, p. 143.

it is clear that Sleeping Beauty's credence in heads on Sunday should be 1/2. But if we place all of the observer-moments in one reference class, we get the following result:

$$\mathrm{C}(heads|Sunday) = \frac{\mathrm{C}(Sunday|heads)\mathrm{C}(heads)}{\mathrm{C}(Sunday|heads)\mathrm{C}(heads) + \mathrm{C}(Sunday|tails)\mathrm{C}(tails)}$$

$$= \frac{\frac{1}{3} \times \frac{1}{2}}{\frac{1}{3} \times \frac{1}{2} + \frac{1}{4} \times \frac{1}{2}} = \frac{\frac{1}{3}}{\frac{1}{3} + \frac{1}{4}} = \frac{\frac{1}{3}}{\frac{7}{12}} = \frac{4}{7}$$

There is a natural way to generalise the preliminary reference class definition to time, which is to treat observer-moments like observers. In other words, any two observer-moments σ and τ belong to the same preliminary reference class if and only if for any world w under consideration, either σ and τ both exist in w or neither σ nor τ exists in w.

Although we have not yet said anything about what would make it means for two possible observer-moments in two worlds to be the same observer-moment, it is clear how this would apply to the Sleeping Beauty case: the two possible observer-moments (i.e. in the heads-world and the tails-world) on Sunday are the same observer-moment; the two possible observer-moments upon awakening on Monday are the same observer-moment; the two possible observer-moments upon being told that it is Monday are the same observer-moment; and the possible observer-moment upon awakening on Tuesday in the tails-world does not exist in the heads-world.

We generally get desirable results about credences if we let every preliminary reference class to consist only of the same observer-moment in different worlds. This makes intuitive sense. Although it is not in any literal sense true, in self-locating reasoning we reason as though we could have been another observer-moment. If I know what time it is now, then I also know that the time that it is now could not have been any other time. When my total evidential state is consistent with multiple times, then my reference class will include only observer-moments that exist at those times. If we were to allow observer-moments at times at which we know are not the present to be part of our reference class, we would get absurd results; for example, the observation that it is whatever time it is now could confirm the hypothesis that I will die sooner rather than later. My reference class must therefore exclude observer-moments that I know are not in the present.

Let Ω_1 be the preliminary reference class of observers-moments on Sunday, let Ω_2 be the preliminary reference class of observer-moments on Monday upon awakening, let Ω_3 be the preliminary reference class of observer-moments upon being told that it is Monday and let Ω_4 be the preliminary reference class of observers-moments on Tuesday. Let '*Sunday*' be Sleeping Beauty's total evidential state on Sunday, let '*awakening*' be her total evidential state upon awakening and let '*Monday*' be her total evidential state when she is told that it is Monday. Then, *Sunday* is consistent only with Ω_1, *awakening* is consistent with Ω_2 and Ω_4 and *Monday* is consistent only with Ω_3. If we let $\Omega_5 = \Omega_2 \cup \Omega_4$, then Ω_1 is Sleeping Beauty's reference class on Sunday, Ω_5 is her reference class upon awakening and Ω_3 is her reference class upon being told that it is Monday. This reference class partition gives us the following results about Sleeping Beauty's credences.

$$\mathrm{C}(heads|Sunday) = \frac{\sum_{\sigma \in \Omega_{heads} \cap \Omega_{Sunday}} \frac{\mathrm{C}(heads)}{|\Omega_1 \cap \Omega(heads)|}}{\sum_{\sigma \in \Omega_{Sunday}} \frac{\mathrm{C}(w_\sigma)}{|\Omega_1 \cap \Omega(w_\sigma)|}}$$

$$= \frac{\sum_{\sigma \in \Omega_{heads} \cap \Omega_{Sunday}} \frac{\mathrm{C}(heads)}{|\Omega_1 \cap \Omega(heads)|}}{\sum_{\sigma \in \Omega_{heads} \cap \Omega_{Sunday}} \frac{\mathrm{C}(heads)}{|\Omega_1 \cap \Omega(heads)|} + \sum_{\sigma \in \Omega_{tails} \cap \Omega_{Sunday}} \frac{\mathrm{C}(tails)}{|\Omega_1 \cap \Omega(tails)|}}$$

$$= \frac{\frac{1}{1}\mathrm{C}(heads)}{\frac{1}{1}\mathrm{C}(heads) + \frac{1}{1}\mathrm{C}(tails)} = \frac{1}{2}$$

$$C(heads|awakening) = \frac{\sum_{\sigma\in\Omega_{heads}\cap\Omega_{awakening}} \frac{C(heads)}{|\Omega_5 \cap \Omega(heads)|}}{\sum_{\sigma\in\Omega_{awakening}} \frac{C(w_\sigma)}{|\Omega_5 \cap \Omega(w_\sigma)|}}$$

$$= \frac{\sum_{\sigma\in\Omega_{heads}\cap\Omega_{awakening}} \frac{C(heads)}{|\Omega_5 \cap \Omega(heads)|}}{\sum_{\sigma\in\Omega_{heads}\cap\Omega_{awakening}} \frac{C(heads)}{|\Omega_5 \cap \Omega(heads)|} + \sum_{\sigma\in\Omega_{tails}\cap\Omega_{awakening}} \frac{C(tails)}{|\Omega_5 \cap \Omega(tails)|}}$$

$$= \frac{\frac{1}{1}C(heads)}{\frac{1}{1}C(heads) + \frac{2}{2}C(tails)} = \frac{1}{2}$$

$$C(heads|Monday) = \frac{\sum_{\sigma\in\Omega_{heads}\cap\Omega_{Monday}} \frac{C(heads)}{|\Omega_3 \cap \Omega(heads)|}}{\sum_{\sigma\in\Omega_{Monday}} \frac{C(w_\sigma)}{|\Omega_3 \cap \Omega(w_\sigma)|}}$$

$$= \frac{\sum_{\sigma\in\Omega_{heads}\cap\Omega_{Monday}} \frac{C(heads)}{|\Omega_3 \cap \Omega(heads)|}}{\sum_{\sigma\in\Omega_{heads}\cap\Omega_{Monday}} \frac{C(heads)}{|\Omega_3 \cap \Omega(heads)|} + \sum_{\sigma\in\Omega_{tails}\cap\Omega_{Monday}} \frac{C(tails)}{|\Omega_3 \cap \Omega(tails)|}}$$

$$= \frac{\frac{1}{1}C(heads)}{\frac{1}{1}C(heads) + \frac{1}{1}C(tails)} = \frac{1}{2}$$

What about Sleeping Beauty's credence that it is Monday upon awakening? Recall that Ω_2 is the preliminary reference class of observer-moments on Monday and that it is a subset of the reference class Ω_5 of observer-moments upon awakening. Then, we obtain the following result that Sleeping Beauty's credence that it is Monday upon awakening is 3/4.

$$C(\Omega_2|awakening) = \frac{\sum_{\sigma\in\Omega_2\cap\Omega_{awakening}} \frac{C(w_\sigma)}{|\Omega_5 \cap \Omega(w_\sigma)|}}{\sum_{\sigma\in\Omega_{awakening}} \frac{C(w_\sigma)}{|\Omega_5 \cap \Omega(w_\sigma)|}} = \frac{\frac{1}{1}C(heads) + \frac{1}{2}C(tails)}{\frac{1}{1}C(heads) + \frac{2}{2}C(heads)} = \frac{\left(\frac{3}{2}\right)}{2}$$

$$= \frac{3}{4}$$

Hence, if we place all observer-moments during the same time in the same reference class (which is necessary in order to avoid absurd results), we obtain the result that Sleeping Beauty's credence in heads should be 1/2 throughout. Clearly, her credence in heads should be

1/2 on Sunday. When she awakens, she does not obtain any evidence for or against heads because she could not have entered any total evidential state other than the one she finds herself in, so her credence in heads remains 1/2. Also when she awakens, she reasons that if the outcome was heads, then it is equally likely that this is Monday and Tuesday. However, when she learns that it is Monday, her credence in heads does not update to 3/4, but stays at 1/2. That is because the fact that this particular awakening occurred is causally independent of the outcome of the coin toss. Regardless of whether the coin is tossed before or after the Monday awakening, the causal process that leads up to the Monday awakening is the same. Hence, the credence in heads stays at 1/2, violating Bayesian conditionalisation.

Hence, my reference class definition, when applied to observer-moments, implies that Sleeping Beauty's credence in heads is 1/2 throughout.

Manley's objections

In 'On being a random sample', David Manley argues in favour of a different kind of theory of self-locating credences. Unlike either Arntzenius and Dorr's theory or Bostrom's theory, which do not weigh worlds with a large number of observers more heavily than they weigh smaller worlds, Manley's theory weighs the prior credence in a world by how many observers it has. It will be instructive to discuss some of Manley's arguments for his view and against the type of view that Arntzenius and Dorr's theory and Bostrom's theory fall under, because some of these arguments will contain potential objections to my view.

Manley's theory consists of two principles: Worldmate Sampling and Frequency. Let us outline each of these principles.

Let us first outline Worldmate Sampling. Let E denote an observer-moment's total evidential state and let p_E denote the posterior credence function of an observer-moment that is in total evidential state E. Now, for any world W that contains at least one observer-moment that is in total evidential state E, let n_W(E) be the number of observer-moments in W with total evidential state E and let n_W(E&H) be the number of observer-moments in W that are in total evidential state E and that H is true of, where H is some self-locating or non-self-locating proposition. Then, Worldmate Sampling states the following:

$$p_E(\mathrm{H}|\mathrm{W}) = \frac{n_W(\mathrm{E\&H})}{n_w(\mathrm{E})}$$ [22]

As it happens, I agree with Worldmate Sampling. Worldmate Sampling is implied by my reference class definition, which becomes obvious if we translate it into Bostrom's framework. As before, let E be a total evidential state, let W be a possible world and let H be a proposition. Since E is a total evidential state, my reference class definition implies that any all observer-moments in total evidential state E in world W belong to the same reference class. Call this reference class Ω. Now, the equivalent of the expression $p_E(\mathrm{H}|\mathrm{W})$ in Bostrom's framework is $P_\alpha(\mathrm{H}|\mathrm{E\&W})$. Hence, we obtain the following:

$$P_\alpha(\mathrm{H}|\mathrm{E\&W}) = \frac{1}{\gamma}\sum_{\sigma\in\Omega_\mathrm{H}\cap\Omega_\mathrm{E}}\frac{P_\alpha(\mathrm{W})}{|\Omega\cap\Omega(\mathrm{W})|}, \gamma = \sum_{\sigma\in\Omega_\mathrm{E}}\frac{P_\alpha(\mathrm{W})}{|\Omega\cap\Omega(\mathrm{W})|}$$

Hence:

$$P_\alpha(\mathrm{H}|\mathrm{E\&W}) = \frac{\sum_{\sigma\in\Omega_\mathrm{H}\cap\Omega_\mathrm{E}}\frac{P_\alpha(\mathrm{W})}{|\Omega\cap\Omega(\mathrm{W})|}}{\sum_{\sigma\in\Omega_\mathrm{E}}\frac{P_\alpha(\mathrm{W})}{|\Omega\cap\Omega(\mathrm{W})|}} = \frac{|\Omega_\mathrm{H}\cap\Omega_\mathrm{E}|}{|\Omega_\mathrm{E}|}$$

This is clearly equivalent to Worldmate Sampling.

Worldmate Sampling is not enough on its own to give us a theory of self-locating reasoning. This is because Worldmate Sampling is a principle of what Manley calls 'inward' reasoning. Inward reasoning means inferring self-locating propositions from non-self-locating propositions. Since H in the above equation may be a self-locating proposition while W, being a possible world, is definitely a non-self-locating proposition, Worldmate Sampling is clearly

[22] Manley, p. 6.

only a principle for inward reasoning. It therefore has to be complemented with a principle for 'outward' reasoning, which involves inferring non-self-locating propositions from self-locating propositions.

Manley's favoured principle for outward reasoning is called Frequency. The overall shape of Manley's argument for Frequency is the following. He considers three candidate principles for outward reasoning: Invariance, Frequency and Typicality. He then argues in favour of Frequency by comparing it with Invariance and Typicality, and arguing that Frequency is superior to both. Let us now state these three principles.

Let us start with Invariance. Again, let E denote an observer-moment's total evidential state, and let p_E denote the posterior credence function of an observer-moment that is in total evidential state E. Then, let E' be the strongest non-self-locating proposition entailed by E (i.e. E' is the set of worlds in which at least one observer-moment is in total evidential state E). Then, Invariance states that the following equation holds for any proposition H:

$$p_E(\mathrm{H}) = p(\mathrm{H}|\mathrm{E}')$$ [23]

Intuitively, what Invariance says is that being in total evidential state E allows us to draw no stronger conclusion than that we are in a world that contains an observer that is in total evidential state E. For example, suppose that E is the self-locating proposition that I am in cell #1 in the Incubator thought-experiment. If E' is the strongest non-self-locating proposition that follows from E, then E' is the proposition that there is an observer in cell #1. When the observer in cell #1 leaves their cell, their total evidential state tells them that they were in cell #1. Invariance, then, says that the observer should respond to this evidence by conditionalizing their prior credence on the proposition that there is (or was) an observer in cell #1. However, the observer in cell #1 already knew this before they left their cell. Invariance, then, says that the credences of the observer in cell #1 should not change when they leave their cell and

[23] Manley, p. 10.

observe that they were in cell #1, because they have learned no new non-self-locating propositions.

Manley rejects Invariance in favour of Frequency. I will not recount Manley's arguments against Invariance because they are not instructive for our purposes. Our purpose here is to find potential objections to my theory from Manley's arguments. But since Invariance is so different from what I propose, Manley's arguments against Invariance are not also potential objections to my theory.

Let us now consider Manley's favoured principle for outward reasoning, Frequency. First, we must define the function *n*. If H is any proposition, *p* is a prior credence function, W is a variable representing worlds and n_W is the number of observer-moments in W that H is true of, then:

$$n(\mathrm{H}) = \sum_{W} p(\mathrm{W}) n_W(\mathrm{H})$$ [24]

Clearly, *n*(H) is the expectation of n_W(H) weighted by the prior credence function *p*. Now, if p_E is again the posterior credence function of an observer-moment in total evidential state E, and H is any proposition, then Frequency states the following:

$$p_E(\mathrm{H}) = \frac{n(\mathrm{E\&H})}{n(\mathrm{E})}$$ [25]

This means that the posterior credence in H of an observer-moment in total evidential state E should be equal to the expected number of observer-moments that are in total evidential state E and are also H, divided by the expected number of observer-moments that are in total evidential state E. To see what this means, let us consider the original Incubator case again. Let E be the total evidential state of being inside a cell and let H be the proposition that heads is the outcome of the coin toss. Then, *n*(E&H) = 50, since this is the expectation of the number of

[24] Manley, p. 20.
[25] Manley, p. 20.

observers in the heads-world who are inside a cell, and n(E) = 101/2, since this is the expectation of the number of observers inside a cell in the incubator. Hence, Frequency says that an observer in the Incubator should have a posterior credence of 100/101 in heads, since there are 100 more observers with the evidential state of being inside a cell in the heads-world than in the tails-world.

Finally, let us consider Typicality, which Manley argues against. First, we must define the function *f*. As before, *p* is the prior credence function and n_W the number of observer-moments in W that a certain proposition is true of, and *I* is the set of inhabited worlds.

$$f(\mathrm{H}) = \sum_{W \in I} p(\mathrm{W}) \frac{n_W(\mathrm{H})}{n_W(all)}$$[26]

Manley uses $n_W(all)$ to mean just the total number of observers in *W*; '*all*' can be thought of as expressing a tautology, and therefore true of any observer in any world *W*. *I* has to be the set of inhabited worlds because if it contained uninhabited worlds, then $n_W(all)$ would be equal to zero for some worlds *W* in *I*. *f* is the expectation of the proportion of *H* observers over all observers in worlds, weighted by the .

Now, if p_E is the posterior credence function of an observer-moment in total evidential state E and H is any proposition, then Typicality says the following:

$$p_E(\mathrm{H}) = \frac{f(\mathrm{E\&H})}{f(\mathrm{E})}$$[27]

f(H) is the expectation of the proportion of all observer-moments that H is true of. p_E(H) is the expectation of the proportion of observer-moments of whom E and H are true, divided by the expectation of the proportion of observer-moments that E is true of. Hence, Typicality says the same thing as Arntzenius and Dorr's Proportion.

[26] Manley, p. 22.
[27] Manley, p. 22.

As I have said, the overall shape of Manley's argument is to argue that Frequency is preferable to both Typicality and Invariance, and that Frequency is therefore likely to be a correct principle for self-locating reasoning. Obviously, I disagree with this argument—Frequency, Typicality and Invariance are not the only three options. Although I have argued against Typicality (this being equivalent to Proportion), Frequency is even further away from my theory. Therefore, it will be instructive to consider Manley's arguments for Frequency and against Typicality also as arguments against my theory.

Now, let us consider Manley's arguments against Typicality and for Frequency.

Manley makes three arguments against Typicality and for Frequency. The first argument is based on two thought-experiments, *Lights* and *Twins*, which support Frequency and go against Typicality according to Manley. The Lights thought-experiment is the following.

> **Lights:** A fair coin is tossed. If the outcome is heads, then one observer that sees light is created. If the outcome is tails, then three observers are created, two of which see light and one of which sees dark. If an observer knows the setup and sees light, what should its posterior credence in heads be?[28]

The Twins thought-experiment is the following.

> **Twins:** A fair coin is tossed. If the outcome is heads, then two observers are created. If the outcome is tails, then four observers are created. Every observer then meets one of the other observers, with the meetings randomly arranged in the case of tails.

Manley says that Typicality involves reasoning as though one were a haecceity that was certain to be instantiated, whereas Frequency involves reasoning as though one were a haecceity that was not certain to be instantiated, but was more likely to be instantiated the

[28] Manley, p. 18.

more observers the world contains.[29] Now, why does Manley think Lights and Twins support Frequency and go against Typicality?

Let us first consider Lights. Suppose that the outcome of the coin toss is tails. Then, two of the observers that are created will see lights, and will therefore be uncertain about the outcome of the coin toss. Since, according to Manley, Typicality involves reasoning as though one were a haecceity that was certain to be instantiated, in the event of tails there will be two observers who reason as though they were both guaranteed to be instantiated—but this cannot be true since, in the event of heads, at most one would have been instantiated, since heads would have produced only one observer.[30]

However, this is not right. Typicality, and similar ways of self-locating reasoning, do not involve reasoning as though one were guaranteed to be embodied. There is an alternative way of illustrating reasoning in the style of Typicality that does not involve reasoning as though one was guaranteed to be instantiated. We can illustrate this alternative way through the example of Lights. There are four haecceities: *s*, *t*, *u* and *v*. In the event of heads, *s* is instantiated. In the event of tails, *t*, *u* and *v* are instantiated in such a way that *t* and *u* see lights and *v* does not. Since the objective chance of heads is 1/2, the objective chance of each of *s*, *t*, *u* and *v* being instantiated is 1/2. Hence, the fact that I find that I am instantiated should not shift my credence in heads away from 1/2, since the objective chance of any haecceity being instantiated was 1/2. So, if we imagine that I find that I am instantiated in Lights but have not yet opened my eyes and hence do not yet know if I am seeing lights or darkness, my credence in heads is initially 1/2—whichever haecceity I am, the objective chance of my instantiation is 1/2, so my mere existence cannot be evidence. Then, since the objective chance of heads was 1/2, I decide that my credence at this point that I am *s* is 1/2, while my credence, by the

[29] Manley, pp. 24–25.
[30] Manley, p. 25.

principle of indifference, that I am any of *t*, *u* or *v* is 1/6. When I open my eyes and see lights, I update my credence function on these priors so that my credence in the proposition that I am *s* becomes 3/5 and that I am either of *t* or *u* becomes 1/5.

Hence, ways of self-locating reasoning that are similar to Typicality, such as Proportion and the Observation Equation, do not involve reasoning as though I am a haecceity that was guaranteed to be instantiated. Instead, they involve reasoning as though I am trying to find out which haecceity I am based on my evidence.

Now, let us consider what Manley says about the thought-experiment Twins. According to Manley, a Typicalist will think about Twins in the following way. The haecceity that I am was certain to be instantiated regardless of the outcome of the coin toss. Suppose I then meet another observer, and call him 'Phil'. If I should reason that I was certain to be instantiated, then Phil should also reason in the same way. In that case, both Phil and I exist both in the heads-world and in the tails-world. In that case, I was certain to meet Phil if the outcome of the coin toss was heads (since Phil and I are the only two relevant observers), but if the outcome of the coin toss was tails then the chance of me meeting Phil would be only 1/3, since I could have been matched up with one of the other two observers instead. Hence, meeting Phil should raise my credence in heads. But this is absurd.[31]

Again, however, the more natural way of reasoning that still gives Typicalist results does not involve positing haecceities that are guaranteed to be instantiated. Instead, it involves positing haecceities that exist only in one outcome. I would reconstruct Twins in the following way. If the coin lands heads, haecceities *s* and *t* are instantiated; if the coin lands tails, haecceities *u*, *v*, *w* and *x* are instantiated. The objective chance of any given haecceities of these being instantiated is 1/2. I know that if I am *s* or *t*, then Phil is the other, and if I am one of *u*, *v*, *w* or *x* then Phil is one of the other three; Phil also only knows that if he is *s* or *t*, then I

[31] Manley, p. 25.

am the other, and that if he is one of *u*, *v*, *w* or *x* then I am one of the other three. But I do not know which of these haecceities I am. I know that the outcome of the coin toss being heads is logically equivalent to me being *s* or *t*, and by the principle of indifference I assign a credence of 1/4 to me being *s* and to me being *t*, and a credence of 1/8 to me being *u*, and to me being *v*, and to me being *w* and to me being *x*. Likewise, I assign a credence of 1/4 to Phil being *s* and to Phil being *t*, and so on. So, again, Typicality does not involve reasoning as though I am a haecceity that was certain to be instantiated. It involves reasoning as though I do not know which haecceity I am, and assigning prior credences in me being various haecceities in accordance with the principle of indifference.

Now, let us consider Manley's second objection. Manley's second objection is that 'it is Frequency, rather than Typicality, that properly handles certain cases where one's guaranteed existence is actually built into the protocol'. He illustrates this with the following example.

> *Gametes*: There are two sperm-egg pairs, A and B, and a fair coin is tossed. In the event of heads, only A will be incubated, and A will see light. In the event of tails, both A and B will be incubated, and A will see light while B will see dark.[32]

The idea here is that the individual resulting from A exists both in the heads-world and the tails-world, while the individual resulting from B exists only in the tails-world. To put it in counterfactual terms, if I see light and the coin landed heads, then I would have seen light had the coin landed tails; and if I see light and the coin landed tails, then I would have seen light had the coin landed heads. Now, Manley says that Frequency is better able to handle a case like this than Typicality. The desired result is that credence in heads conditional on seeing light is 1/2, while credence in tails conditional on seeing dark is 0. This is what Frequency implies, while Typicality implies that one's credence in heads conditional on seeing light should be 2/3, and in heads conditional on seeing dark is 0.

[32] Manley, p. 26.

Manley extrapolates from this that when one's guaranteed existence is built into the protocol, that Frequency gives the intuitively correct results while Typicality gives intuitively wrong results. In cases like Adam and Eve, some observers' guaranteed existence is also built into the scenario, namely, Adam's and Eve's, and it is cases like this that cause problems for Arntzenius and Dorr, who would agree with Typicality. However, Typicality and Frequency are not the only two possible options, and reference classes exist precisely to handle cases like these, where 'some observers' guaranteed existence is built into the protocol', as Manley puts it. If some observer is guaranteed to exist, and another observer is not, then they will belong to two separate reference classes according to my theory. In a case like the one that Manley is thinking of, with a set of observers that are guaranteed to exist regardless of the outcome of a chance event and another set whose existence possibly depends on the outcome of the chance event, the two sets of observers belong to two separate reference classes within any world. Credence in any outcome conditional on being guaranteed to exist is equal to the prior credence in that outcome, according to this reference class definition.

Finally, let us consider Manley's third objection. Manley's third objection is that Typicality depends on a clear definition of what counts as an observer, whereas Frequency does not. This is because Frequency only counts observers with the same total evidential state. On the other hand, when we use Typicality or similar principles, we may need to know whether, for example, a chimpanzee counts as an observer. Manley illustrates this with the thought-experiment *Dog*.

> **Dog:** A fair coin is tossed. If it lands heads, a human is created. If it lands tails, a human and a dog are created. I am created as a human; what should my posterior credence in heads be?[33]

[33] Manley, pp. 27–28.

Frequency implies that my posterior credence in heads should be 1/2, regardless of whether dogs are considered observers. On the other hand, if we accept Typicality then the result depends on whether we consider dogs to be observers. If we do, then the posterior credence in heads given that the observer is a human, should be 2/3; otherwise, it should be 1/2.

However, the fact that Typicality and similar principles assume objective criteria for what kinds of entities count as observers is not in itself a problem for Typicality and similar principles provided that there is no reason to think that the notion of observerhood is intrinsically incoherent or problematic in some way. As far as I can see, there is no reason to think this. There may be vague cases of observer, such as, according to Manley, dogs. Such vague cases of observers may make the principle difficult to apply in some cases. However, the mere fact that a principle is difficult to apply in some cases does not make the principle itself incorrect, and there are many reasons besides whether something counts as an observer that might make a principle difficult to apply.

In order to improve Typicality, or any similar principle that depends on a clear definition of what an observer is, such as Proportion or the Observation Equation, the notion of observerhood could be analysed in more detail. We already have an intuitive sense of what it means to be an observer, which is why it is possible to state a principle like Proportion. If we look at entities in the world, in most cases we can clearly divide them into non-observers and observers. Bottles and chairs are not observers, while humans, hypothetical intelligent lizards, sufficiently intelligent robots, and so on, are observers. Hence, a definition of observerhood would reduce the incidence of vague cases and confirm the intuitions that we already have about observers. A natural candidate for a criterion of observerhood might perhaps be phenomenal consciousness. This does not entail any metaphysical commitments, such as that consciousness is irreducible to physics, but only that it objectively exists in the world.

Taking Manley's example of dogs specifically, my intuition is that they clearly are. The fact that they are less intelligent than us and do not have certain abilities that we do, does not exclude them from being observers. Manley's example of dogs seems problematic for two reasons. The first reason is that dogs cannot think about self-location. But we should not think that all observers in our reference class have to be able to reason about self-location. The second reason is that Manley's thought-experiment Dog does not make it sound like the human's existence depends on the outcome of the coin toss. It sounds as though the coin toss only decides whether a dog will be created in addition to a human that will be created independently of the outcome of the coin toss. Then, the human and the dog are in separate reference classes when reasoning about the coin toss.

Finally, let us consider Manley's last remark in his comparison of Typicality and Frequency. Earlier in this essay, we have already described the thought-experiment of the Presumptuous Philosopher, who concludes that one physical theory is almost certainly correct because it predicts a much larger number of observers, which Bostrom and Arntzenius and Dorr both use to argue against Frequentist intuitions. The Presumptuous Philosopher expresses the thought that mere existence should not be taken to be evidence that there are more observers as opposed to fewer; that an observer's prior credence in a hypothesis should be equal to its credence in the hypothesis conditional on the observer's own existence. Manley's response to Presumption is a similar thought-experiment that intuitively seems to favour Frequency over Typicality.

> **Presumption 2:** 'As in *Presumption* except the relevant theories are T3, which says there are a trillion non-green subjects in the universe and a trillion trillion green subjects, and T4, which says there are a trillion of each.

‘The typicalist, having noticed that she’s non-green, will declare it completely unnecessary to test these theories empirically, because T4 is a trillion times more likely than T3. This seems pretty presumptuous as well.’[34]

I think that we have the intuition that the Typicalist in the second version of Presumption is presumptuous for the following reason: we do not know with certainty that the green subjects and the non-green subjects in the T3-world belong in the same reference class. Recall that two observers belong to the same reference class if and only if they either both exist or neither of them exists. This typically depends on whether there is a single causal process that creates a group of observers or several distinct causal processes such that it is possible for some of them to occur while others fail to occur—in the latter case, the observers created by distinct causal processes belong to different reference classes.

In Presumption 2, it would be presumptuous to not to test the theories empirically because we do not know, if we are in the T3-world, whether the non-green subjects are in the same reference class as the green subjects. If the theory T3 says that the same causal process created both the green subjects and the non-green subjects, then they are in the same reference class. If, on the other hand, T3 says that the existence of those green subjects that exist is independent of whether T3 or T4 is true—or, in other words, if the same green subjects exist in the T3-world and the T4-world—then the green subjects are in separate reference classes. If the thought-experiment Presumption 2 were to specify this, then there would be no difficulty for my theory, and it would give the intuitively right results in each case. If the green and the non-green subjects are in different reference classes, then our observation that we are green would not give us any evidence for T3 or for T4, and we would have to test which theory is true empirically. If, on the other hand, T3 definitively tells us that the same causal process created the green and the non-green subjects, so that any given green observer exists only in

[34] Manley, p. 32.

the T3-world or the T4-world but not both, then observing that we are green would in fact be very strong evidence against T3 and in favour of T4, and it would not be presumptuous to update our credences just on the basis of self-locating evidence. So, the intuition of presumptuousness in Presumption 2 comes from the fact that the case is under-specified.

What should we say about the original version of Presumption? A proponent of Frequency, like Manley, may have the opposite intuition in the case of Presumption; they may think that in a case like this, we should in fact treat our existence as evidence for a theory that predicts a larger number of observers. By comparison, before other planets were discovered (but after we knew that the Earth is round), someone might have treated the observation that we are on Earth as evidence for the hypothesis that we live in a universe with many planets. Or, if I walk into a forest and see just one fox, I can treat this as evidence that the forest contains many (as opposed to a small number of) foxes. Why not also say that when we observe ourselves, we treat this as evidence that there are many observers in the universe?

If we want to strengthen the intuition for Typicality and against Frequency, we can consider the original Incubator case once more. Recall that in the original Incubator, 100 observers are created in the event of heads and 1 observer is created in the event of tails. Typicalism and my theory tell us that when an observer awakes in the incubator, its credence in heads should be 1/2, while Frequentism says it should be 100/101. Manley says that Typicalism says that an observer reasons as though they were a haecceity that was certain to be instantiated regardless of the outcome of the coin toss. However, Typicalism does not depend on or require us to reason this way. In the Incubator case, an observer who awakens in a cell does not think ‘I would have the same experiences even if the coin had landed opposite of how it has actually landed’. The observer knows that if the coin toss had landed opposite of how it has actually landed, then he would not exist. Rather, when the observer awakens in the incubator he does not receive any evidence for or against heads or tails. He cannot reason that his existence is evidence for anything, because he does not know which observer he is. To say that he is more likely to be one of the observers in the heads-world because there are more of them is to beg the question for Frequentism. Therefore, the observer's credence in heads is 1/2 upon waking.

We can say something similar about Presumption. The physicists in the thought-experiment must have reached some values for the objective probabilities of T1 and of T2. Whatever evidence and theories they used in reaching these objective probabilities must have also included the observation that we exist as observers. Therefore, that we exist is old evidence and we cannot conditionalize on it again. Since, by hypothesis, observers' typical experiences do not differ between T1 and T2, our credences in T1 and T2 conditional on our experiences cannot be different from our prior credences in T1 and T2. This is similar to the situation that an observer in the incubator is in. By hypothesis, they know the set-up, so they know that the objective probability of heads is ½, and when they awaken they do not receive new evidence that would allow them to shift their credence in heads away from 1/2. When the observer exits their cell and observes that they were in cell #1, they are able to rule out that they are any of the possible observers in cells #2–100. Crucially, the observer does not reason that they could have been in a cell other than #1, and so they do not reason as though they are a haecceity that was guaranteed to be instantiated. In fact, they do not reason as though they were any kind of haecceity. Rather, they start out not knowing which actual observer they are, and narrow down their identity on the basis of self-locating evidence.

Hence, my theory does not involve reasoning as though haecceities exist, or that one could have existed in another world. It involves not knowing one's own identity at the beginning, and narrowing it down on the basis of self-locating evidence. On the other hand, Frequency does involve reasoning as though haecceities exist. Specifically, it involves reasoning as though one is a haecceity that potentially exists in every world and has a chance of being instantiated proportional to the number of observers in that world. The claim that haecceities exist, or that an observer σ could have been another observer τ to whom σ is not identical, is dubious and unnecessary. Hence, this consideration favours my theory and goes against Manley's.

Conclusion

In this essay, I have outlined a theory of self-locating credences based on Bostrom's Observation Equation. First, I outlined Arntzenius and Dorr's theory of self-locating credences. I argued that their theory cannot be correct, because of the strongly counter-intuitive results that it gives for cases like Adam & Eve and UN++. Second, I outlined Nick Bostrom's theory. I argued that Bostrom's theory is not a complete theory because without a reference class definition, the theory does not tell us anything about what observers' credences should be, and Bostrom does not provide any kind of reference class definition. Following this, I developed and explained my own reference class definition. I began with a preliminary definition according to which the reference class of an observer *s* is the set of observers *t* such that for any world *w*, either *s* and *t* both exist in *w* or neither *s* nor *t* exists in *w*. This definition is inadequate because in general, an observer's evidential state may be consistent with more than one preliminary reference class according to this definition. I then gave a general reference class definition, according to which an observer's reference class is the union of preliminary reference classes that the observer's total evidential state is consistent with. Finally, I discussed the Sleeping Beauty problem from the point of view of my reference class definition. I concluded that Sleeping Beauty should have a credence in heads of 1/2 throughout—upon awakening on Monday because the reference class of the Monday awakening includes only other observer-moments in the same total evidential state, which exist in both worlds, and upon being told that it is Monday because the Monday awakenings are causally independent of the coin toss. Lastly, I defended my theory from possible objections from Manley.

Bibliography

Bostrom, Nick (2002). *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. New York: Routledge.

Dorr, Cian and Arntzenius, Frank (2014). 'What to Expect in an Infinite World'. Unpublished Manuscript.

Dorr, Cian and Arntzenius, Frank (2017). 'Self-locating Priors and Cosmological Measures' in Khalil Chamcham, John Barrow, Simon Saunders & Joe Silk (eds.), *The Philosophy of Cosmology*. (pp. 396–428). Cambridge: Cambridge University Press. Accessed online on 15 February at https://philpapers.org/archive/DORSPA-4.pdf

Elga, Adam (2000). 'Self-locating belief and the Sleeping Beauty problem'. *Analysis*, 60.2, pp. 143–147.

Manley, David (n. d.). 'On being a random sample'. Unpublished Manuscript. Accessed online on 15 February at https://static1.squarespace.com/static/55d3621de4b07a4744ce4a23/t/55f91ca4e4b0fa2519ad7a73/1442389156509/OBARS.pdf

Lewis, David (1980). 'A Subjectivist's Guide to Objective Chance' in Richard Jeffrey (ed.), *Studies in Inductive Logic and Probability, Vol. 2*. (pp. 263–293). Berkeley: University of California Press.