THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

DE SE BELIEFS AND CENTRED UNCERTAINTY

SILVIA MILANO

A thesis submitted to the Department of Philosophy, Logic and Scientific Method of the London School of Economics for the degree of Doctor of Philosophy, London, January 2018

DECLARATION

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I declare that my thesis consists of 59,375 words.

Silvia Milano

ABSTRACT

What kind of thing do you believe when you believe that you are in a certain place, that it is a certain time, and that you are a certain individual? What happens if you get lost, or lose track of the time? Can you ever be unsure of your own identity? These are the kind of questions considered in my thesis. Beliefs about where, when and who you are are what are called in the literature *de se*, or self-locating beliefs. This thesis examines how we can represent *de se* beliefs, and how we can reason about *de se* uncertainty.

In the first part of the thesis, I present and motivate a specific account of the content of *de se* belief, based on the one given by David Lewis. On this account, the content of *de se* beliefs are centred propositions. I defend this view against a rival account, put forward by Robert Stalnaker, according to whom the content of *de se* beliefs are ordinary (non-centred) propositions.

In the second part of the thesis, I explore how we can reason probabilistically about *de se* uncertainty. I start by defining probabilities over centred propositions, and investigate what probabilities mean in this context. As it turns out, all the main interpretations of probability can be extended to centred propositions. The only trouble seems to arise for the Bayesian principle of updating via conditionalization. After giving a diagnosis of the problem, I offer a solution by formulating a natural extension of conditionalization, which I argue preserves the essential features of Bayesian reasoning.

In the final chapter, I apply my view and show that it leads to a natural resolution of a puzzle (known as the Sleeping Beauty problem) that is generally taken to be a test case for any account of centred updating.

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Christian List, for inspiring me to write this thesis and for his constant support and encouragement. No matter how busy he was, Christian always found the time when I had an issue to discuss. This thesis would have been substantially poorer without his detailed and constructive feedback, and I could not have asked for a better supervisor. I would also like to thank Anna Mahtani for being a fantastic second supervisor. I would not have been able to finish this thesis without her help, insightful comments and criticism. She exemplifies for me the playful love of knowledge that always drew me to philosophy.

I had the great fortune of being part of a wonderful philosophical community at the LSE. I am grateful to the Philosophy department, to Richard Bradley and the AHRC Managing Severe Uncertainty project for the support that I received, and to Bryan Roberts, Katie Steele, Alex Marcoci, Aron Vallinder, Paul Daniell, Johannes Himmelreich, James Nguyen, Goreti Faria, Chris Marshall, Tom Rowe, Deren Olgun, Kamilla Buchter and Todd Karhu for stimulating discussions on material related to this thesis at various points during these four years. I would also like to thank my fellow PhD students for creating such a friendly environment. Spending time with you made me a better philosopher and a happier person.

Finally, I would like to thank Paul Froment for his love and support through the good and the bad times, especially during the last few months of writing this thesis. Very special thanks go to my sisters, Alessandra and Angela, and to my parents, Nadia and Marcello, for encouraging me to pursue what I am passionate about.

CONTENTS

LIST OF TABLES

'But if I'm not the same, the next question is, Who in the world am I? Ah, that's the great puzzle!' And she began thinking over all the children she knew that were of the same age as herself, to see if she could have been changed for any of them.

– L. Carroll, *Alice's Adventures in Wonderland*

1

INTRODUCTION

If you have ever been lost, unsure of where you are or what time it is, then you know what it is like to lack some piece of self-locating, or *de se* information. This is information about where, when and who you are in the world. I am fairly certain that this kind of situation is common (and occasionally troubling) enough that it will be obvious why, intuitively, one should care about *self-locating* information. But why should philosophers, and specifically epistemologists, be interested in *de se* beliefs and *de se* uncertainty? What is special about them?

First of all, *de se* beliefs are extremely pervasive. Chances are that as you are reading this text, you will most definitely have many such beliefs of your own, as I do myself. For example, I'm aware of who I am (Silvia), I know that I'm in London, that it is a Monday night in December and that as I type, it is already dark outside. I'm not exactly sure of the time, but I believe that if I wanted to check it right now, I could turn on the screen of my phone.

Moreover, it seems to me that all these beliefs about one's identity, spatial and temporal location play an important role in shaping one's experience, and the way one reasons about the world. Some philosophers, following John Perry (1979), have argued that *de se* beliefs are essential to explain our very sense of

agency (see Chapter 3 below). If I stop writing to make myself a cup of tea, for instance, at least part of the explanation for this action is that I believe that I myself would like a cup of tea now, and that I also believe that it is in my power to make one.

What's more, as I will argue in Chapter 6 *de se* information is essential to understanding the link between observations and evidence, which is a fundamental component of Bayesian reasoning. A careful study of the content of *de se* beliefs will lead us to reflect on some key features of Bayesian rationality, solving some quirky probability puzzles along the way (see Chapter 7).

The overall plan of my thesis will be the following. In the first part of the thesis, I will present and motivate a specific account of the content of *de se* belief, based on the one given by David Lewis. On this account, the content of *de se* beliefs are centred propositions. I defend this view against a rival account, put forward by Robert Stalnaker, according to whom the content of *de se* beliefs are ordinary (non-centred) propositions. In the second part of the thesis, I explore how we can reason probabilistically about *de se* uncertainty, and advance a solution to a diachronic puzzle that *de se* beliefs appear to raise for Bayesian reasoning.

1.1 GENERAL PLAN

Despite their pervasiveness and relevance, *de se* beliefs do not fit very well with a standard account of the content of beliefs that many philosophers, including epistemologists, subscribe to. On this view, the content of beliefs are propositions, understood as sets of possible worlds. For example, suppose that I believe that London is a city of eight million: the proposition that I believe contains all the possible worlds where London is a city of eight million, and

does not contain all the possible worlds where London has fewer or more than eight million inhabitants. Here, a possible world is intuitively just a specific way in which the world could objectively be. The most famous proponent of this view, David Lewis (1986), thought that all possible worlds exist – not just the actual world. But we don't really need to take a stand regarding the metaphysical status of possible worlds here, and in the rest of this thesis I will just assume that we can use them to provide a semantic framework for the content of beliefs, without settling whether they do in fact exist in a metaphysical sense. However, possible worlds are too coarse grained to capture the content of *de se* beliefs, such as for example my current *de se* belief that it is now Monday. We will need some extra machinery to pick out a location, time, an agent within a possible world.

The extra machinery that is needed to perform this task are *centres*. In Chapter 2, I present and discuss some examples of *de se* beliefs and introduce centred worlds as a formal device to model *de se* information. Intuitively, a centred world is just a pair of a possible world and a centre within it. I then review several applications of the centred worlds framework, and consider interpretations of the framework that have been proposed in the literature, concluding that, even though which interpretation we choose may depend on the specific application, centred worlds provide the right sort of formal framework to study *de se* uncertainty.

Following on this discussion, Chapter 3 gives an analysis of a famous example originally introduced by Perry, arguing that two intuitively plausible readings of this example correspond to two modes of reasoning about *de se* information, which I call the *cartographer* and the *pathfinder* modes. I argue that while these two modes of reasoning can often appear mixed together in practice, they correspond to opposite 'directions' of inference: on the cartographer mode, one uses *de se* information to reconstruct what the world is objectively

like. On the pathfinder mode, on the other hand, one tries to use objective information about the world to figure out one's own location. I then present two rival accounts of *de se* uncertainty that are formulated within the centred worlds framework, due to Robert Stalnaker (2008) and to David Lewis (1979), and show that only the latter is compatible with both modes of reasoning.

As I will argue in Chapter 4, Lewis's account presents several advantages. Besides being compatible with both modes of reasoning about *de se* beliefs that I identify in Chapter 3, it permits a natural application of probabilities. On Lewis's account, which is the one that I subscribe to, the contents of *de se* beliefs are simply centred propositions, i.e. sets of centred worlds. This represents a natural refinement of the possible worlds framework, and allows a straightforward identification of belief states with their contents. Stalnaker's alternative account, on the other hand, divorces *de se* beliefs from their content. According to Stalnaker, *de se* beliefs should be modelled as sets of centred worlds, but the content of *de se* beliefs always correspond to ordinary, non-centred propositions. I present and criticise Stalnaker's account in Chapter 4.

We have seen that *de se* beliefs are both pervasive and relevant. Can we assign probabilities to the contents of uncertain *de se* beliefs? I consider this in Chapter 5, where I start by defining probabilities over centred propositions, and investigate what probabilities mean in this context. I find that defining probabilities over sets of centred worlds requires no specific technical modifications to the theory of probability, and that all the main interpretations of probability can be extended to centred propositions. The only trouble seems to arise for the Bayesian principle of updating via conditionalisation.

In light of this, in Chapter 6 I focus on the diachronic puzzle that centred probabilities raise for Bayesian reasoning. It is now generally accepted in the literature that *de se* beliefs are not compatible with the standard Bayesian principle

of updating via conditionalisation (see Titelbaum, 2016b). I review the current literature on the topic, focussing in particular on the account of updating as communication put forward by Sarah Moss (2012), discussing the theoretical assumptions about rationality that underpin Moss's account. I then present an alternative account, which allows a natural extension of conditionalisation to centred events, and argue that it is compatible with the essential features of Bayesian reasoning.

Finally, Chapter 7 applies my view to the Sleeping Beauty problem. This problem has attracted considerable attention in the literature as a paradigmatic example of how self-locating uncertainty 'creates havoc' for standard Bayesian principles of conditionalisation and reflection, and it is also thought to raise serious issues for diachronic Dutch Book arguments. I show that, contrary to the consensus view, it is possible to represent the Sleeping Beauty problem within a standard Bayesian framework. Once the problem is correctly represented, the solution satisfies all the standard Bayesian principles, including conditionalisation and reflection, and is immune from Dutch Book arguments. Moreover, the solution does not make any appeal to the Restricted Principle of Indifference that is generally accepted in the literature on self-locating uncertainty, which, I argue, is incompatible with the principles of Bayesian reasoning.

1.2 LOOKING FORWARD

As I have outlined in this introduction, *de se* beliefs are an interesting research subject for epistemologist. They are pervasive, relevant and raise interesting puzzles for one of the most widely accepted interpretations of probability. As I take it, the main achievement of this thesis is to show that reasoning about

de se beliefs is in fact compatible with the fundamental principles of Bayesian reasoning. This surprising conclusion opens up interesting further questions, some of which will be outlined in Chapter 6.

2

CENTRED WORLDS

In this chapter, I introduce centred worlds and motivate their use to represent self-locating uncertainty. I first present some examples of self-locating uncertainty in §2.1, explaining some of the reasons why we should be concerned with this type of uncertainty. In §2.2, I motivate the choice of centred worlds as the means to represent self-locating uncertainty. §2.3 reviews the key applications of centred worlds in philosophy and some other related disciplines. Finally, §2.4 presents the accounts of centred worlds that are currently present in the literature. These accounts differ from each other in how they individuate centres.

2.1 SELF-LOCATING UNCERTAINTY

Limited agents like us are often ignorant or uncertain about various features of the world and our place within it. On the one hand, we often lack complete information about what the world is like. For example, imagine that Emma has inadvertently spilled some wine on her dress, and is trying to use water to remove the stain, seemingly unaware of the fact that water does not remove

wine stains. Or, to take a different example, imagine that a coin is about to be tossed, and you don't know yet what the result is going to be.

In both cases, the relevant uncertainty seems to be about some objective features of the world – 'what the world is like' – that are independent of the particular perspective or location from which they are entertained as possibilities. Either water removes wine stains, or it does not, and this fact is independent of whether Emma is aware of it or not. Similarly, either the coin lands Heads, or it lands Tails (assuming that it is tossed under normal conditions). Which outcome actually occurs is a feature of the world that is independent of whether and by whom it is observed.

Typically, we can express possibilities of the kinds illustrated by the two examples above in terms of possible worlds. A possible world corresponds – roughly – to a maximally detailed description of the world, specifying all the facts that hold true within it. For example, whether wine stains can be removed with water will be one of the facts that are true at some worlds, but false at others. So this possibility divides possible worlds in two classes, the ones where it is true that water removes wine stains, and the ones where it is false. Accordingly, the actual world must fall in either one of these two classes.

Similarly, either the coin toss comes up Heads, or it comes up Tails. Each outcome corresponds to a different way that the world might turn out to be like. Either the actual world is a Heads-world, or it is a Tails-world, and it is so independently of who might be able to observe the result of the toss. Perhaps, if the coin is yet to be tossed, it may be yet indeterminate or impossible to know in advance which outcome takes place, until the toss has taken place.[1] But once the toss is performed and that bit of information is settled, only one of the two

1 This will depend on the interpretation of probability that one accepts. See Chapter 5 for a discussion.

possibilities can be true, independently of the method, timing or perspective of an observation.

The two examples that I just presented illustrate cases of uncertainty about what the world is like. At times, however, agents could also lack information about their own position within the world, their own identity or the role that they occupy. Following a standard usage, I will call uncertainty that loosely falls under any of these types *self-locating uncertainty*.

Self-locating uncertainty does not seem to be straightforwardly tied to lack of information or ignorance about what the world is like. Even in cases where agents possess all the relevant information about the objective features of the world, they may still be unable to identify the position that they occupy, as the following examples will help to illustrate.

Example 1. Missing coordinates After a violent storm, Tom's ship is lost at sea. Luckily, he has a detailed digital map of the region where he is sailing, but the GPS system is damaged and does not display Tom's current location. High at sea, around noon, the surroundings lack any specific element that would enable Tom to pin down exactly where he is.

Example 2. Hidden display Ann forgot to take out her watch before going into the shower, and as a result the screen is now completely fogged. The watch itself is not damaged, but she won't be able to read the time until the fog clears away, which will probably take a few hours. In the meantime, Ann can't be sure of what time exactly is being displayed.

The next example is due to John Perry (1979):

Example 3. The messy shopper 'I once followed a trail of sugar on a supermarket floor, pushing my cart down the aisle on one side of a tall counter and

back the aisle on the other, seeking the shopper with the torn sack to tell him he was making a mess. With each trip around the counter, the trail became thicker. But I seemed unable to catch up. Finally it dawned on me. I was the shopper I was trying to catch.' (Perry (1979), p. 3.)

In the first example, Tom is uncertain of his current spatial location, as he does not know his coordinates. Let us suppose now that he considers the possibility 'I am at point (x, y, z)'. Clearly, as he considers it, this possibility might be true or false. But it is not true or false *eternally*, because Tom's position could (and likely will) change over time, so Tom could be located at (x, y, z) at some times, but not at others.

Similarly, in the second example Ann is uncertain about the time currently displayed on her watch. She already knows what times the display is going to show and in which order, but as she's lost the ability to look at the display she does not know what it is showing right now. Suppose that she considers the possibility that the time displayed now is 7:05 a.m. This possibility is clearly going to be either true or false at any time that Ann entertains it. But, again, it is not true or false *eternally*,[2] because the time displayed by the watch changes regularly.

The third example has received much attention in the literature on self-locating uncertainty, and it is the one with which John Perry is often credited to have introduced what is known as the problem of the *essential indexical*. An essential indexical is so called because it expresses some information about *who* a certain individual is, which it would not be possible to express by making recourse only to non-indexical terms. At his final moment of realisation, the protagonist in the messy shopper example appears to learn something new, namely that he is the messy shopper. Moreover, this seems also to be reflected in the way that

2 At least on a first thoughts – but I will return to this issue later on, and I will discuss this point in more detail in Chapters 3 and 4.

his behaviour is likely to change – for instance, we might now expect that he will stop and check his cart for a torn sack of sugar. However, Perry argues that it is very difficult to pin down what the content of this new piece of information exactly is. Perry's point is that this content cannot be expressed simply in terms of new objective information about what the world is like. Instead, he claims that the first personal pronoun *I* in the sentence 'I am the messy shopper', which expresses what the protagonist realises, is an essential indexical.

According to Perry (1979)[3] the messy shopper does not learn any new information about what the world is like when he realises who he is. For instance, Perry was aware from the start that a messy shopper was leaving a sugar tray, and that he himself was walking around the counter just as the messy shopper was, and so on. What he did not know, however, is which possible individual he was. Consider again the sentence 'I am the messy shopper', which is what Perry comes to believe in the example. Given that Perry already knew that there was a messy shopper around, he knew from the start that 'I am the messy shopper' would be a true sentence, when evaluated from the perspective of the correct individual (the one who happened to be the messy shopper). As it turns out, Perry himself is the messy shopper, so 'I am the messy shopper' is true for him. But this sentence is not true or false absolutely. It is true from Perry's perspective, but false from the perspective of Sally, another shopper in the same store.

Whereas the missing coordinates and the hidden display examples highlight how some sentences can take different truth values depending on the spatial or temporal location from which they are evaluated, Perry's messy shopper example illustrates a case where a sentence can take different truth values depending on the identity or the role occupied by the agent from whose perspec-

3 See also Liao (2012) and Lewis (1979). I will return on the example of the messy shopper in Chapter 3, where I will also use it to motivate a more detailed discussion of different modes of reasoning with self-locating uncertainty.

tive the sentence is evaluated. All three are examples of self-locating or *de se* uncertainty.

2.2 CENTRED WORLDS

As the three examples presented in the previous section suggest, sentences like 'I am at point (x, y, z)', 'the display shows 7:05 a.m.' or 'I am the messy shopper' are not true or false absolutely, but only with respect to the specific location (in space or time), or the role occupied by the relevant individual.

Assuming that the map is an adequate representation of the objective geographical features of the world, Tom knows exactly what the world is like. However, he is still unable to locate *himself* within it. Given Tom's circumstances, the sentence 'I am at point (x, y, z)' could be either true or false, but its truth value seems subject to change. If Tom's ship moves, then the sentence might become true, if it was previously false, or *vice versa*.

Similarly, Ann knows exactly the sequence of symbols that the fogged display of her watch is going to produce. However, she is unable to say precisely which time the screen is indicating right now. Assuming that Ann knows that at some point the watch is going to display 7:05 a.m., Ann knows that the sentence 'the display shows 7:05 a.m.' has been, is, or will eventually be true. However, like in the previous case, the truth value of this sentence seems to change as different times are successively displayed on the screen.

In both examples, there are some pieces of information – which, following a standard practice I will called *uncentred* information – that correspond to features of the world that do not change with respect to the agent's location. In the missing coordinates example, this corresponds to the information contained

in Tom's map, while in the hidden display example the uncentred information corresponds to what Ann knows about the functioning of the watch and the sequence of times that it is going to display.

Alongside the uncentred information, the agents in both examples also have some *centred* information, that is information relative to the location that they currently occupy. In the missing coordinates example, Tom's centred information includes all the points on the map that appear compatible with his surroundings. In the hidden display example, the centred information includes all the time points compatible with what the watch might be displaying. Since Ann cannot see the display and she has no other way of synchronising her present beliefs with the time time shown by the watch, her centred information contains all the time points that are compatible with her current rough estimation of what time it is.

It is a bit less straightforward to distinguish the centred and uncentred components of the information held by the agent in the messy shopper example. Prior to the realisation that he is the messy shopper, John Perry has access to uncentred information about his surroundings, including the fact of being aware of the presence of a messy shopper who is leaving a trail of sugar on the floor, and of a shopper who is trying to catch the messy shopper in order to stop him. He also has some centred information. He knows, for instance, that he himself is the shopper who is trying to catch the messy shopper, whose identity is, for all that Perry knows, uncertain.

As the previous examples highlight, the truth value of self-locating sentences can change with respect to the context at which they are uttered. There are two different morals that we could draw from this observation. On the one hand, we could say that sentences like 'I am at point (x, y, z)' or 'the display shows 7:05 a.m.' have a fixed meaning, but the truth value of self-locating sentences

like these is determined by the context of utterance and therefore is subject to change. On the other hand, we could maintain instead that sentences like 'I am at point (x, y, z)' or 'the display shows 7:05 a.m.' do not have a fixed meaning, but rather that the meaning is fixed by the context at which they are uttered. In other words, if we choose this route, self-locating sentences are in a sense incomplete, as in order to fix their meaning it is necessary to add in the context of utterance. On the resulting contextualist picture, the truth value of any self-locating utterance is fixed.

I have written so far as though the truth value of an utterance or thought can change over time. This may not be accurate, and there are many complex issues here concerning utterances, truth and indexicals. Here I have skated over these issues for the purposes of this introduction, but I will return to an investigation of these issues in chapters 3 and 4, where I will analyse more closely different proposals to capture the semantic content of *de se* expressions.

Following a substantial literature originating from Lewis (1979), we can represent predicaments such as the ones illustrated by these three examples using *centred worlds*. Intuitively, a centred world (w, c) is a pair of a possible world w and a centre c within it. The possible world component w corresponds to a specific way that the world could be like, encoding all the uncentred information about that world, while the centre c picks out a specific location or individual within w, encoding all the centred information relative to that perspective on w.

Possible worlds can be used to represent objective (or non-self-locating) possibilities, but they are not fine-grained enough to express self-locating possibilities. In order to extend our account of propositions to centred worlds, we will need to introduce a few distinctions. Firstly, we should amend the general definition of a proposition. On this revised account, a proposition p (centred or

uncentred) is just a set of centred worlds. Similarly as before, we will say that a proposition p is true at a centred world (w, c) if and only if (w, c) belongs to p.

Secondly, we should make a distinction between centred and uncentred propositions. I will take an *uncentred* proposition p to be a set of centred worlds satisfying the following condition: for any pair of centred worlds (w, c) and (w, c') that differ only with respect to the centre, p contains (w, c) if and only if it also contains (w, c'). Conversely, any proposition p that does not satisfy this condition will be called centred.

For example, let us take p to be the proposition 'Easter Island is 3.700km far from Lima'. As he is lost at sea, Tom might believe that p is true, based on the information on his map. His belief in p does not depend on where he believes is actual location to be. Believing p does not give Tom any information regarding his own location. Since p is true at all possible locations that Tom could occupy, it contains all centred worlds (w, c) such that w is the actual world (corresponding to the information on the map) and c is a centre within w. Conversely, let q be the proposition 'I am at location (x,y,z)'. q is true at some, but not all the locations within w that Tom could occupy, so there are some pairs of centred worlds (w, c) and (w, c') such that q is true at (w, c), but false at (w, c').

2.2.1 *Some preliminary objections and replies*

Before moving forward in the discussion, in this section I consider some preliminary objections to the introduction of a centred worlds framework to reason about self-locating uncertainty.

Specifically, I can anticipate the following two main objections:

1. We don't need *centres*: Possible worlds are sufficient to capture all relevant possibilities. We don't need self-locating possibilities: they are reducible to standard, non-self-locating possibilities.

2. We don't need *centred worlds*: Instead, we should look at alternative accounts of propositions that are not formulated in terms of possible worlds, and which would be able to handle the cases of self-locating uncertainty.

If correct, the first objection, which is a version of the reductionist or deflationist thesis about self-locating uncertainty, represents a serious challenge to centred worlds. If all cases of self-locating uncertainty, such as the examples illustrated in the previous section, can ultimately be reduced to instances of regular, uncentred and non self-locating uncertainty, this would seem to imply that the whole exercise of introducing centred worlds is not only pointless, but also wrong-headed. Introducing centred worlds would be pointless because since there are not really different kinds of uncertainty demanding a different treatment, regular possible worlds are all we need. It would also be wrong-headed, because it would introduce distinctions where no distinctions ought to be made. This is an issue in so far as it might sanction incorrect ways of reasoning about uncertainty, depending on the framing of the problem.

I don't believe that this objection is ultimately successful, because not all centred propositions might be reducible to uncentred ones, and therefore self-locating uncertainty really presents some special issues that could not be solved without the introduction of centres. I will discuss this point in more detail in Chapters 3 and 4. But independently of whether the deflationist thesis is ultimately true, we also have some independent reason to introduce centred worlds. This is because even if it were always possible to reduce self-locating possibilities to non-self-locating ones, it will often be the case that the reduc-

tion comes at the cost of significantly increasing the complexity of the representation. This is not at all an unusual problem for reductivist theories. For example, one could make an analogy with reductionist thesis that are sometimes discussed with respect to natural or social sciences. It is sometimes argued, for instance, that psychology can in principle be entirely reduced to the physical processes going on inside the brain. Without entering in the merit of this discussion, it is – I think – reasonable to expect that even if the reductionist thesis were true, the simpler and most effective way to investigate psychological facts might still be to use the tools offered by psychology.

The second objection is motivated by the observation that there are a class of well-known puzzles about reference that have proved very difficult to solve for many of the more popular theories of proposition, and self-locating beliefs seem to share a lot of features with the puzzles in this category. Building on this point, some authors including Magidor (2015) and Cappelen and Dever (2013) have argued that all the motivations generally invoked for introducing centred worlds are actually simply versions of well-known puzzles about reference involving possible worlds, where the issue is not necessarily that of self-location or essentially indexical terms. They do not deny (as the deflationist does) that there might be essential indexicals or self-locating uncertainty that are impossible to express on the possible world account of propositions. However the conclusion that they would draw from this is not that we should refine the possible world account of propositions by introducing centres, but rather that we should abandon it altogether in favour of a different account (I will return on this point again briefly in Chapter 4, §4.1 and §4.1.1).

Centred worlds indeed may inherit some metaphysical issues from possible worlds. However, these issues should be considered separately, and the account that I will develop in further chapters does not presuppose a specific metaphysical view regarding what centred worlds really are. Moreover, as will

become clearer in the discussion (see in particular Chapter 5), centred worlds provide an ideal framework to reason about self-locating uncertainty, since probabilities can naturally be defined over them. Other accounts of propositions, on the other hand, do not allow for the same natural algebraic structure. For example, the accounts considered in Magidor (2015), do not allow the same straightforward extension of probabilities to self-locating possibilities.

2.3 APPLICATIONS

In this section, I review some key applications of centred worlds in philosophy and some neighbouring disciplines. Centred worlds are generally introduced as a modelling device to represent centred or self-locating possibilities. In certain areas, such as for example in the philosophy of action or in decision theory, centred worlds frameworks have become almost the norm.

One of the main reasons for having centred worlds is that they are useful to explain actions. In Perry's messy shopper case, for example, the new centred belief acquired by Perry ('*I* am the messy shopper!') explains why he stops to check his cart, instead of continuing to walk around the counter. Similarly, in the Hidden Display example, Ann's centred belief that it might be 7:05am can explain why she starts to prepare some coffee. Another main reason to introduce centred worlds is to give a better account of the content of mental attitudes such as beliefs and desires, or aesthetic and moral judgements that appear to be sensitive to the perspective from which they are formulated. In all these cases, centred worlds allow for a more fine-grained representation of the content of attitudes. In the Missing Coordinates case, for example, the content of Tom's belief state can be represented as the set of all the centred worlds that he takes to be compatible with his current surroundings. Since all these

alternative locations coincide with respect to the uncentred component, a simple possible worlds representation of Tom's belief state would not be enough to express his uncertainty regarding his own location. Centred worlds can also be used to model the contents of desires. For instance, in the Hidden Display example Ann might hope that it is now 7:05am, so that she still has the time to make some coffee before she leaves for work.

Centred worlds also received some applications in computer science. For example, in a series of articles Joseph Halpern and other collaborators have used centred worlds to model the local states of asynchronous distributed systems Halpern (2004). In these application, possible worlds correspond to the possible runs of an 'experiment', while centred worlds are used to model the 'information state' of agents that receive limited information about the experiment, concerning the state of the system at a given (but not always known) point in time. In linguistics and the philosophy of language, centred worlds are used as a foundational object to study the semantics of indexicals, to capture contextual information, or to model self-centred talk (see Stojanovic (2016)). Dilip Ninan uses a multi-centred version of a centred worlds framework to model the semantics of counterfactual attitudes and linguistic expressions. In the philosophy of mind, centred worlds figure in the two-dimensional semantics put forward by David Chalmers (see Chalmers (2006)).

Berit Brogaard's perspectivalism, a theory that she applied to solve problems in ethics Brogaard (2012), aesthetics and epistemology, takes centred worlds as a foundational entity:

> Perspectivalism is a semantic theory according to which the contents of utterances and mental states (perhaps of a particular kind) have a truth value only relative to a particular perspective (or standard) determined by the context of the speaker, assessor, or bearer

> of the mental state. I have defended this view for epistemic terms, moral terms and predicates of personal taste elsewhere. (Brogaard, 2010)

A competing relativist approach, which also uses centred worlds as foundational entities, has been developed by Andy Egan. His relativist theory has received various applications in ethics (Egan, 2012), philosophy of language and aesthetics (Egan et al., 2005; Egan, 2010). In all these cases, Egan argues that centred worlds allow us to better understand both the sense in which aesthetic and moral properties can be objective, and how different individuals might rationally disagree about aesthetic and moral judgements. Egan has also applied his theory to metaphysics, where he has argued that centred worlds can shed light on the nature of secondary properties (such as colour, for instance), which – unlike the primary properties – seem to be observer-dependent (see Egan, 2006a,b).

As the last application mentioned indicates, on Egan's brand of relativism centred worlds can be taken to correspond to metaphysical possibilities. Not all applications, however, can be plausibly interpreted in this way. Most of the applications mentioned above in decision theory, philosophy of action and philosophy of language simply use centred worlds as a tool to model behaviour, beliefs and other mental attitudes.

Based on this observation, Liao (2012) identifies two families of approaches to centred worlds in the philosophical literature. On the one hand, the *epistemological* approaches take centred worlds as a way to model and reason about self-locating uncertainty. These approaches take centred worlds to represent epistemic possibilities, the objects of beliefs, desires or other mental attitudes, without necessarily endorsing specific claims about the metaphysics. On the

other hand, *metaphysical* approaches view centred worlds as fundamental entities.

As Liao notes, these two different approaches might generate different, possibly incompatible applications. For example, it seems plausible that the theoretical entities that play the role of centred worlds in Egan's treatment of secondary qualities would not be the right candidates to represent local states of distributed system in Halpern and Fagin's framework. In other words, even if the applications mentioned in the previous section all use a similar formal notion of centred worlds, we cannot assume that the precise interpretation of this notion will be consistent across all applications.

This brings me to another observation. As evidenced by the overview in this section, centred worlds provide a useful formal tool across several applications. But we should not mistake the formalism for something more substantive. Perhaps different interpretations of the formalism can be appropriate, depending on the intended applications for which it is introduced.

2.4 WHAT ARE CENTRES?

In this section, I move on to review the main accounts of centred worlds that are currently present in the literature. The presentation follows closely the one in Liao (2012), who identifies the four accounts discussed below.

In his paper, Liao is primarily interested in the question of what centres *are*, which can naturally be interpreted as a metaphysical question. The way he approaches answering this question, however, sees him analysing how different accounts 'pick out' or identify centres. On Liao's view, the centres are therefore whatever theoretical entities are 'picked out' or identified by the right account.

A problem for Liao's strategy, however, is that (as I pointed out in the previous section) different accounts might be suitable for different intended applications, and yet be incompatible among each other. If this is the case, then there might be no metaphysical entities underpinning the role of centres in each account. Moreover, it might be difficult to decide on which account is 'correct', independently of the intended applications, unless one has a prior understanding of the metaphysical notion of what a centre actually 'is', which could not be delivered by Liao's strategy.

A possible way around this first worry is given by Liao's proposed identification of centres with 'possible individuals', corresponding to the primitive identification account discussed in §2.4.4. This offers a direct way to access what centres are, insofar as we have some understanding of what possible individuals are. Unfortunately, the question of what possible individuals are is also a difficult one, as the criteria for personal identity are far from being uncontroversial. As a result, for Liao, '[t]he question of what centres are is intimately related to the question of how possible individuals are individuated' (Liao (2012), p. 7).

In light of this problem, Liao's choice to equate possible individuals and centres does not help directly to answer his central question ('what are centres?'). But it may still deliver some other results. For example, our intuitive understanding of what possible individuals are might possibly place restrictions on how centres should be individuated.

As will be discussed below, my view is that requiring the identification of individuals and centres may be both unnecessary and unhelpful. It seems to add complications to the problem of identifying centres (which now must also accommodate common intuitions about personal identity), without bringing a definite advantage in terms of answering Liao's central question of what centres are.

Moreover, the identification of centres with possible individuals appears to severely restrict the possible applications of the framework. For example, we might want to use centred worlds to model fictional scenarios, or to imagine how things would look from a perspective that no-one occupies, when reasoning counterfactually. A possible example might be Einstein's famous thought experiments in his exposition of general relativity theory. The ideal observer who travels at the speed of light might not be a metaphysically possible individual, for instance, but this does not make Einstein's examples intuitively unintelligible. We could represent the perspective of Einstein's ideal observer as a set of centred worlds, and yet identifying each centre with a possible individual would be incorrect.

So, to summarise: I agree with Liao on the functionalist maxim, or, as we might put it in a slogan, that 'a centre is what a centre does'. However, this maxim does not automatically place constraints on what entities might play the role of centres in the context of different applications. Unlike Liao, I am not too interested in the metaphysical question of what centres really are, across all applications, as I believe it may be more fruitful to concentrate on what centred worlds frameworks can achieve in different intended applications. Moreover, I don't find it convincing that we can take the functionalist maxim to provide an answer to the metaphysical question of what centres are. This question might not even be meaningful, if the intended applications of centres worlds are sufficiently far apart as not to allow for a unified treatment.

While Liao is mainly concerned with the metaphysical question of what centres are, the question that I am more interested in is epistemological: how can one know the centre that one occupies? A review of the accounts of centred worlds that are present in the current literature will help me to map out the conceptual space within which the centred worlds framework is situated, its canonical interpretations, and limitations.

2.4.1 *The Quinean account*

According to what is known as the *Quinean account*, centred worlds are a pair (w, c) of a possible world (w) and a centre (c), where c is identified as follows:

> Given a system of coordinates R, a centre c is a point in R.

Quine is often credited as the first philosopher to have introduced centred worlds as a refinement of possible worlds to express self-locating features. In Quine (1968), he proposed centred worlds as a device to represent the content of propositional attitudes such as beliefs and desires.

To introduce centred worlds, Quine uses a famous example of a cat (who I will name Oscar):

> 'We like to say for instance that the cat wants to get on the roof, or is afraid the dog will hurt him. In so saying we purport to relate the cat perhaps to a state of affairs. The cat wants, or fears, the state of affairs.' (Quine (1968), p. 10).

But what is the state of affairs desired, or feared by Oscar? One possible candidate would be that Oscar desires a state of affairs, or a possible world, where there is a cat on the roof, and fears a state of affairs where a cat is in the clutches of a dog. However, Quine argues, this simple answer is not at all satisfactory. Consider, for example, a state of affairs that contains both a cat on the roof, and another cat in the clutches of a dog. Then according to the simple answer just given, Oscar would both want and fear this state of affairs at the same time. But this is clearly wrong: what Oscar apparently wants is not just a state where there is a cat on the roof, but *he* wants to be that cat.

Quine's proposal to capture this self-locating component of the attitudes we ascribe to Oscar involves the introduction of centred worlds (or, as Quine calls them, centred states of affairs). When we say that Oscar wants to be on the roof, according to Quine, we are ascribing to Oscar a desire whose content is a set of centred worlds, all of them centred on a cat on the roof (and not on a cat in the clutches of a dog).

If we accept Quine's proposal, the possible world component (the w) of a centred world corresponds to the uncentred features of a state of affairs, while c corresponds to a location within w. This location corresponds to a point, individuated by a set of coordinates, relative to a coordinate system that is defined on w. For instance, if w encodes how matter is physically distributed and space-time is Newtonian, a centre can be identified as a point determined by its spatio-temporal coordinates. Individuals, like cats, dogs or people can also be identified by the coordinates of the spatio-temporal region that they occupy. Going back to the example, according to Quine the content of Oscar's desire corresponds to the set of centred worlds that are centred on the region corresponding to the cat on the roof, and excludes all those that are centred on the region corresponding to a cat in a dog's clutches.

The Quinean account can also be used to represent self-locating possibilities in the examples that I presented in the previous section. For instance, the Quinean account can handle the missing coordinates example quite straightforwardly. Suppose that after sailing towards the East for some time, Tom sees a profile on the horizon, that could only correspond to a certain island on the map. From this observation, he comes to believe that he is located at a specific coordinate point x with respect to his map, in other words he locates himself at a centred world that is centred on x.

In more general terms, we can summarise Quine's proposal as follows. On the Quinean account, centres are locations within possible worlds that are individuated by their coordinates. The system of coordinates that is used to identify centres, in turn, is defined on a possible world w.

Quine argues that the choice of coordinate system is to some degree arbitrary. For example, it seems that we could vary the unit of measurement, or the origin, without changing the nature of what is measured. Moreover, in principle there seems to be no requirement that the coordinates defined on different possible worlds should be similar in some relevant respects. For example, world w might be a bi-dimensional world, whereas w' might contain four dimensions. However, in practice, we might be interested in cases of centred uncertainty where the centred worlds that are considered possible belong to worlds with similar systems of coordinates. This is the case in all the preceding examples: in the Missing Coordinates case, all the centred worlds have centres that are identified by a triple of spatial coordinates. In the Hidden Display case, the relevant centres are identified by different time points. And finally, in the Messy Shopper case, centres could be identified by spatio-temporal coordinates defined on a similar space.

An important thing to note is that the Quinean account does not explicitly identify centres with possible individuals. Instead, the possible individuals are picked out relative to their coordinates, or the possible centres that they might happen to occupy. To illustrate this point, we can look at how the Quinean account treats Perry's messy shopper example. What Perry comes to believe in that case, upon realising that he is the messy shopper, is that he and the messy shopper are in fact the same person. The Quinean account can handle this case by saying that what Perry comes to believe is that he himself and the messy shopper are co-located. In other words, after the realisation, Perry

comes to believe that for any possible centred world (w, c), Perry is located at (w, c) if an only if the messy shopper is located at (w, c).

It might be debatable whether the Quinean treatment of the messy shopper case is satisfactory. In one way, it might seem to just miss the point: learning that he is co-located with the messy shopper does not automatically seem to imply that Perry *is* the messy shopper. For starters, if the coordinates used to individuate centres are coarse grained enough, we might generally expect several individuals to be co-located. For example, if the coordinates used to identify the centre are coarse-grained enough to pick out regions of space corresponding to, say, the area of an entire neighbourhood, then we would normally expect different agents to be co-located at the same centre. The Quinean account could try to circumvent this issue by drawing on the modal content of what Perry learns. Saying that Perry is located at a centred world (w, c) if and only if the messy shopper is also located at (w, c), whatever the level of grain of the system of coordinates used to identify the centre would, I think, fix the issue. In other words, if we allow the coordinates to vary arbitrarily (including sufficiently fine-grained coordinates), we should normally reach a point where the sets of centred worlds at which different agents are located are different.

Liao, however, following Lewis (1979), raises a further objection against this strategy. To illustrate this objection, Liao describes a scenario where the exact same region of space that is occupied by John Perry (and the messy shopper), is also occupied by a conscious Ghost. Ghost is distinct from Perry (and is not a messy shopper!), and yet the content of the belief that Perry comes to believe, namely that he is co-located with the messy shopper, is the same as the content that he is co-located with Ghost.

The case of the Ghost raises a problem for the Quinean account, but only depending on the underlining assumptions that we are willing to make about the

possibility of co-located individuals. There might be some less metaphysically complicated cases that have this feature. For example, we might perhaps construct a virtual reality where multiple individuals are located at the same coordinates. The case I have in mind here could be similar to a computer simulation where some set of characters always occupy the same location. Arguably, the relevant coordinate system to individuate each character is the virtual coordinate system in which they live.

In any case, even if Liao's objection weren't successful, it reveals that the Quinean account has some difficulty to accommodate cases, such as the Messy Shopper case (and possibly the virtual reality example), where the uncertainty is primarily about one's own identity. This is because the Quinean account does not have the resources to express essential indexicals (such as the content of 'I am the messy shopper') directly. Instead, it must rely on some background assumptions about the possibility of co-location of individuals. For this reason, while still possible, the Quinean account might be less suitable if our primary aim for introducing centred worlds is to handle this kind of cases.

2.4.2 *The Lewisian account*

On what is known as the Lewisian account, centres are identified as follows:

> A centre c is a pair (i, t) of an individual i and a time t.

The so-called Lewisian account of centres solves the Quinean account's problem with co-location of different individuals by stipulating that centres are identified on the basis of the individuals that occupy them. This solves the problem of co-location, because two different agents (such as Perry and the

Ghost) who might share the same spatio-temporal coordinates would nonetheless be located at different centres.

The individual i is usually taken to be a persisting person or agent, such as John Perry in the messy shopper example, or the cat Oscar in the example from Quine (perhaps stretching the definition of a person a little bit!), while a natural way to specify t is to fix it relative to an external or objective time dimension. For instance, in the hidden display example Ann was uncertain which time her watch is currently displaying. Since she also believes that the watch is accurate, on this version of the Lewisian account we can identify each centre that she considers possible as an ordered pair $c = (t_i, a)$ where a stands for Ann, and t_i is the objective time.

An issue that Liao raises for the Lewisian account is that, just as the Quinean account seems unable to distinguish between co-located individuals, the Lewisian account is unable to distinguish between distinct stages of the same continuing person in cases of time travel. To illustrate, we can again follow Liao to present a modification of the missing coordinates example. Imagine again that at time t Tom sees a profile on the horizon and realises what his present location is. The Lewisian account would express this by saying that the content of Tom's newly acquired belief is a centred world $(w, (t, Tom))$, where t is the present time, and Tom is the continuing person on which the world is centred. But imagine now that in the future, Tom learns to time travel and travels back to the same time at which he was lost at sea after a violent storm. In this scenario, both older-Tom and younger-Tom are present at t, but at different locations. Clearly, the contents of the beliefs of younger- and older-Tom should be different, but on the Lewisian account both correspond to the same set of centred worlds, namely all the worlds (w, c) where the centre is identified by Tom and a time t. This is because both younger- and older-Tom are present at

t and they are both parts of the same continuing person, so both are located at the same centre, namely (t, Tom).

The moral that Liao draws from the time travel example is that the Lewisian account does not specify an adequate way to identify centres. However, time travel is notoriously a source of paradoxes, and it is highly doubtful if we can at all make sense of its notion, given that it is incompatible with our most advanced physical theories. So, a different moral that we could draw from the example recounted by Liao is that it is simply an illustration of the incoherence of the possibility of time travel.

Another potential issue for the Lewisian account, which might make it unsuitable for some applications, is that it may be unable to identify some possible centres. On the Lewisian account, each centre has to correspond to a possible individual. However, there are cases where we might want to identify some centres at which no individual is located. An example of such a case is how we sometimes reason counterfactually about how things would appear from a different perspective than the one that we actually occupy. The Missing Coordinates example helps to illustrate this point. As Tom is lost at sea, he might reason about his surroundings in the following way. He knows from the map that, if his current location was closer to the coast, he would be able to see something on the horizon. Since he does not see anything on the horizon, he concludes that his current location is not close to the coast. For this intuitive way of reasoning to work, we do not need to assume that the alternative locations considered by Tom are all picked out by different individuals, nor do we need to assume that an individual is located at all of them.

Moreover, using individuals to identify centres may be impractical. In Tom's case, for example, we might think that the coordinate system identifies uncountably many points, which intuitively correspond to centres at which Tom

might be located. However, since the temporal coordinate t is the same for all these, according to the Lewisian account they must correspond to different individuals. This means that we need uncountably many individuals to identify all the possible locations for Tom in the Missing Coordinates example. This poses some obvious problem: we wouldn't be able to name all these individuals in any natural language, for instance, and indexing them by their spatial coordinates would just seem to bring us back to the Quinean account.

An advantage of the Lewisian account over the Quinean account is that it gives a more intuitive account of essential indexicals. The content of a statement like 'I am the messy shopper' (believed by John Perry), for example, is represented on the Lewisian account as the set of centred worlds (w, c) such that John Perry is the messy shopper at w, and c corresponds to John Perry and the current time t.

This relative advantage with respect tot he Quinean account is, however, paid for on other accounts. Firstly, it makes the Lewisian account vulnerable to problems about reference, such as Frege's puzzle. If John Perry initially believes that he and the messy shopper are distinct individuals, the worlds centred on John Perry and those centred on the messy shopper are two disjoint sets. But then it is difficult for the Lewisian account to explain how it is that John Perry might come to believe that the two sets of centred words do, in fact, coincide. This change cannot be modelled as a simple case of updating. I will say more about this case in chapter 3.

2.4.3 *The exhaustive set account*

A third account, which Liao calls the *exhaustive set* account, identifies centres through an exhaustive ordered set of properties.[4] In Liao's words:

> The ordered set of [properties] being exhaustive guarantees that there cannot be a possible scenario where two individuals share all of the mentioned [properties] but differ on an unmentioned [property]. Since the ordered set is exhaustive, there cannot be any other [property] to differ on. [...] On the exhaustive set account, possible individuals [*and, hence, centres*] are individuated by all the [properties] they could have. If two possible individuals differ on any [property], they are distinct. (Liao, 2012, p. 311)

To flesh out what an exhaustive set of properties is, we need to provide a list of properties that an individual could (in principle) have. This might already be a difficult task: to do this we may need to say something more about what class of individuals we are looking to, or what properties are. But as it turns out, we can give the following as a general definition that captures the exhaustive set account:

> A centre c is identified by a *maximally consistent* set of properties.

A set of properties P is said to be *consistent* if there is a set of individuals that satisfy all the properties in P. Moreover, P is said to be *maximally consistent* if there is no other property p such that p or its negation could be added to P and P would still be consistent. In other words, all the individuals that satisfy a maximally consistent set of properties could not, by the way this set is constructed, differ on any other property p which is not in the set P.

4 Instead of properties, Liao uses the word *features* to refer to the items used to pick out centres.

As this discussion clearly shows, the fact that P is maximally consistent does not in itself guarantee that it will pick out a unique individual. One way to guarantee that this is the case is by having among the possible properties the identity properties of each object. Since any individual can only be identical with itself, this ensures that each set of maximally consistent properties corresponds to a unique individual.[5]

A drawback, however, is that adding the identity properties makes it difficult to square the exhaustive set account with either of two views of properties. If properties are understood extensionally, i.e. if a property is just the set of all the individuals that have it, then each maximally consistent set of properties is just formally equivalent to an identity property. But this is a problem, if we wanted to use the properties to identify the possible centres, in accordance with Liao's reason for introducing exhaustive sets of properties as a way to pick centres. If, on the other hand, properties are understood intensionally, then it seems unclear that having identity properties could be helpful to pick out centres. In other words, for any individual i, the corresponding identity property of 'being identical to i' seems particularly uninformative: to use the intensionally-understood identity property 'being identical to i' to actually *individuate i* seems just to reverse the order of things.

2.4.4 *The primitive identification account*

The fourth (and final) account of centred worlds identified by Liao differs from the previous three, because it does not specify any way of identifying centres.

5 Liao (see pp. 312-3) discusses identity properties in the primitive identification account, and not the exhaustive set account, but – in light of the definition of a maximally consistent set given above – I think this is wrong.

Instead, it takes centres to be primitives. So, on the primitive identification account, the definition of a centred world is simply the following:

> A centred world (w, c) is an ordered pair of a possible world w and a centre c.

The primitive identification account is the one favoured by Liao, and he also convincingly argues that it might be closer to the account that Lewis originally had in mind.[6] However, adopting the primitive identification account also comes at a price: namely, it makes the individuation of centres somewhat mysterious. Since centres are identified with possible individuals, and the latter are taken to be primitive, this account tells us nothing about how we can individuate centres unless we are also provided with some background information about possible individuals.

Like the Lewisian account, the primitive identification account is committed to the existence of individuals. However, unlike for the Lewisian account, individuals should not be taken to be persisting persons. In almost every other respect, the primitive identification account shares the advantages (and some disadvantages) of the Lewisian account, but presents some additional problems.

The primitive identification account is good at explaining essentially indexical statements, for the same reasons as discussed for the Lewisian account. However, compared to the Lewisian account, the primitive identification account seems to make it more difficult to explain an agent's actions on the basis of her attitudes towards centred possibilities. Let's say, for example that Oscar the cat (O) 'believes' his current situation to be the set $S = \{(w, (O, t_0)), (w', (O, t_0))\}$, where w corresponds to a possible world where Oscar is on the roof at t_1, and w' is a possible world where Oscar is by the dog at t_1. Since at t_0 Oscar is in

6 Liao (2012), p. 313-4. See also Lewis (1986, 1983).

all cases in the same situation and far from the dog, he has no reason to prefer one centred world over the other. But we might say that Oscar 'desires' it to be the case that, at t_1, the centred world at which he is located is $(w,(O,t_1))$. This would explain why Oscar takes the action to jump on the roof. On the primitive identification account, however, Oscar at t_0 and Oscar at t_1 are counted as two distinct individuals,[7] seemingly making it more difficult to see in what sense Oscar could be said to 'desire' a certain state of affairs, or how this could give *him* a reason to act.

Like the Lewisian account, the primitive identification account has some trouble accommodating identity statements, for roughly the same reasons discussed in the previous section. I will also have more to say on this in Chapter 4, §4.1.1.

2.5 CONCLUSION

In this chapter, I have reviewed four different accounts of centred worlds that are present in the literature. As noted in the opening section, different accounts of centred worlds might be suited to different intended application. Something that can be drawn from my discussion is that if the purpose of introducing centred worlds is primarily to express the content of essentially indexical statements, then the Lewisian account, or the primitive identification account, might be the most suitable. Both these accounts, however, make it impossible for distinct individuals to agree on centred contents and have some problems to accommodate identity statements.

The Quinean account, on the other hand, gives an intuitive treatment of agreement (different individuals can agree on centred content, if they share the

7 The same is also true for the two centred worlds corresponding to Oscar at t_0: on the primitive account, each is associated with a distinct individual.

same location relative to the relevant system of coordinates). It also gives a more natural way to express the content of centred counterfactual statements, and, if taken in conjunction with some background assumptions about the possibility of co-location of individuals, it is able to accommodate identity statements. However, the Quinean account does not seem to deliver an intuitively satisfactory account of essential indexicals.

3

TWO MODES OF REASONING

In this chapter, I discuss a famous example due to Perry (1979), and argue that two intuitively plausible readings of this example correspond to two ways in which we naturally reason about *de se* possibilities (in what I call the *cartographer* and in the *pathfinder* modes).

I then present the two currently most widely accepted semantic frameworks to model *de se* expressions and beliefs – namely the one due to Stalnaker (2008) and the so-called 'Lewisian' framework (based on Lewis (1979)) – and show that only the latter is compatible with both modes of reasoning for *de se* uncertainty.

In conclusion, I outline the advantages and disadvantages of both frameworks, noting that if the contextualist account that is at the basis of Stalnaker's framework is correct, then this would rule out the possibility of the pathfinder mode.

3.1 THE CASE OF THE MESSY SHOPPER

The case of the messy shopper is an example, initially introduced by Perry (1979), that is often used to explain why we need centred worlds to express what Perry calls 'essentially indexical' facts, which possible worlds are not fine grained enough to capture.

Here is Perry's original example:

> I once followed a trail of sugar on a supermarket floor, pushing my cart down the aisle on one side of a tall counter and back the aisle on the other, seeking the shopper with the torn sack to tell him he was making a mess. With each trip around the counter, the trail became thicker. But I seemed unable to catch up. Finally it dawned on me. I was the shopper I was trying to catch. (Perry, 1979, p. 3)

The problem is: what is it that Perry learns when he realises that he is the messy shopper? There seem to be two plausible answers:

- He learns something about the world: namely, *which individual is the one who is making a mess.*
- He learns something about himself: namely, *which individual he himself is.*

In the next sections, I will consider each answer in turn. I will look at the problem in the context of the possible (centred) worlds view of propositions. Other accounts of propositions possible (see Magidor (2015), and Perry himself adopts a different account of propositions), but centred worlds have the advantage of allowing a straightforward extension of probabilistic reasoning

to *de se* uncertainty (as will be explained in more detail in Chapter 5). As standard, I take possible worlds to represent different ways the world might be (see Chapter 2). For the purpose of analysing Perry's messy shopper case, we can take centres to correspond to different individual agents within a possible world. Let $W = \{w_1, \ldots, w_n\}$ be the set of all worlds that Perry considers possible, given the information he possesses, and $C = \{a_1, \ldots, a_n\}$ be the set of possible centres. A centred world (w_i, a_j) is an ordered pair of a possible world w_i and a centre, a_j, which specifies an individual agent within w_i. The possible world w_i expresses facts about what the world is like, whereas a_j picks out a specific location, or individual perspective, within w.

3.1.1 *About the world*

Let $W = \{w_1, w_2, \ldots, w_n\}$ be a set of possible worlds compatible with Perry's evidence. For definiteness, we will assume that each w_i specifies who, among a set of possible individuals who exist at that world, is the one causing a mess.

To illustrate, suppose that there are three individuals in each world, who go by the names of Jones (a_J), Smith (a_S) and Perry (a_P), respectively. As he notices the trail of sugar on the floor, Perry comes to believe that someone is making a mess, but he does not know the identity of this individual. His evidence at this point is therefore compatible with the following three possible worlds:

- w_1 = 'a_J is making a mess';
- w_2 = 'a_S is making a mess';
- w_3 = 'a_P is making a mess'.

All the while, Perry is not uncertain about his own identity, so he knows that the actual world must in any case be centred on a_P. Perry does not know whether he is or is not making a mess, because he does not know which of w_1-w_3 is the actual world. We can represent Perry's total belief state before he realises that he is the messy shopper as the set of centred possibilities $B = \{(w_1, a_P), (w_2, a_P), (w_3, a_P)\}$. All the centred worlds contained in this set are centred on the same individual, namely a_P, and differ only with respect to the possible world component. In our scenario, Perry is certain that one of these possibilities corresponds to his present circumstances, but he does not yet know which.

At the point in which he finally realises that he is the messy shopper, we can think that Perry acquires some information that enables him to single out w_3 as the actual world. For example, he might have realised that Perry is the only one who has been walking around the counter in a way that is consistent with the sugar trail left on the floor. And he might, at this point, stop to check his cart for a torn bag of sugar, thereby testing the hypothesis that the messy shopper is indeed John Perry. This way of representing the problem makes Perry's uncertainty to be entirely about the world, so a natural question to ask is whether the centred component is really needed to capture Perry's beliefs in this case. In other words, since we are assuming that Perry is certain about his own identity throughout, why don't we represent Perry's total belief state simply using the set of possible worlds $\{w_1, w_2, w_3\}$?

A standard response to this observation is that specifying a centre might still be necessary if the goal is to explain how beliefs can motivate action. If I have a belief that Silvia will be late for dinner unless she leaves immediately, I will not be motivated to leave, unless I also identify myself as Silvia. Similarly, believing that a_P is making a mess would not in itself explain why Perry stops and checks his cart, unless we add the further premise that Perry identifies himself with a_P.

In other words, upon realising that w_3 is the actual world, Perry is motivated to act in a certain way (e.g., check his cart for a torn bag of sugar) *because* he believes of himself that he is Perry. Without this background belief, which is expressed by the centred component of (w_3, a_P) there seems to be no reason for Perry to act in this particular way.[8]

The sort of background belief that is needed to motivate action (e.g. the belief that I am Silvia, or Perry's belief that he is Perry) is a belief about which individual agent one is. Perry (1979) calls this kind of beliefs *essentially indexical*, because they are usually expressed using sentences containing indexical terms such as 'I', or demonstratives such as 'this' or 'that'. For example, the essentially indexical belief that motivates Perry to check his cart can be expressed using the sentences '*I* am the messy shopper' and '*This* is the messy shopper's cart'. Following an established convention in the literature,[9] I refer to these as *de se* expressions and beliefs. *De se* beliefs are necessary to explain action, but *de se* truths could not be deduced from a purely objective description of the world. For this reason, the specification of centres is needed to express the essentially indexical component of an agent's beliefs.

In this section I've been assuming that in the Perry example what is learnt is just something about the world. Nonetheless, even under this assumption, we can see that there may still be a role for *de se* beliefs. For it seems that Perry must have a background *de se* belief (that he himself is Perry) to explain why what he learns motivates him to check his cart.[10]

8 See also Stalnaker (2016). I will have more to say on this in Chapter 4.
9 See e.g. Ninan (2016), Paul (2017) among others.
10 This observation is also relevant to Stalnaker's account of *de se* beliefs, and I will return on this point in chapter 4.

3.1.2 *About the centre*

Suppose that Perry has been able to work out that w_3 ('a_P is making a mess') is the actual world. In other words, based on his observations, he now knows that a_P – out of the three possible agents, a_P, a_J, and a_S – is the agent who is making a mess. However, a_P, a_J, and a_S name three distinct individuals within the world, but Perry does not know which one *he himself* is.

We can represent Perry's uncertainty using the centred worlds framework, only this time Perry's uncertainty is not about which possible world corresponds to his present circumstances, but rather he is uncertain about which centre he occupies within that world. On this way of reading the messy shopper example, then, Perry's belief state before he realises that he is the messy shopper corresponds to a set of centred possibilities $B = \{(w_3, a_J), (w_3, a_S), (w_3, a_P)\}$. The centred possibilities contained in B coincide with respect to the possible world component (w_3), but differ with respect to the centred component, to reflect the fact that Perry believes it possible that he himself actually identifies with any one of the three agents who are present at w_3. When he finally realises that he himself is the messy shopper, on this reading of the example, Perry is able to eliminate some possible centres. As a result, he identifies (w_3, a_P) as the actual centred world at which he is currently located.

3.2 TWO MODES OF REASONING

The exploration of the two intuitive readings of the messy shopper case reveals that two distinct modes of reasoning may be at play in each. In this section, I will outline these two modes of reasoning, explaining how they relate to each

of the two readings of the messy shopper example, and how they could be generalised to other cases.

3.2.1 *The cartographer mode*

On the first reading of the messy shopper case that I have identified, Perry is seen as inferring something about the world from the observation of his own surroundings. I will refer to this way of reasoning as the *cartographer mode*. To illustrate how this mode operates, we can make an analogy with the job of a cartographer. The cartographer sets out to map the geographical features of a territory about which she does not yet have detailed information. We can imagine she is dropped by helicopter in some unknown location, and to map the territory, she starts making observations from different points. The map that she produces in this process reproduces the features she observed, recording their relative positions, within a specified system of coordinates. The idea here is that the agent, at any given point, is making observations that give her some more information about the world. For example, if from the point where she is dropped, on a hill top, she can see a gorge and a river streaming through it to the North, she now knows that there is a river and a gorge – and not, for example, a field – to the north of the hill where she is standing.

Now the cartographer knows that there is a gorge and a river, but how far are they from the hill? In order to answer this question and record that information on the map, she needs to measure the distance. Taking her current location to be at the origin of the initial observation, she could use some appropriate instruments or move towards the river while keeping track of the distance walked. In this way, as the cartographer moves around exploring the territory, she is able to make more observations. One of the observation points (the

initial one, in this case) is kept fixed as the origin, and the other observation points are identified by their spatial coordinates with respect to the origin.

In this example, there is nothing special about the origin point: in this case, it makes sense for the cartographer to fix it as the initial point at which she starts making her measurements, keeping track of the relative distances from there. But the important information is contained in the objects observed (rivers, gorges, hills) and the relative distances between them. The coordinate points could be re-labelled at a later time, or the origin fixed at a different point, without altering the essential features of the map (Quine, 1968).

The resulting map represents the territory, not just as viewed from a particular point, but in an abstract way. This makes it possible to communicate the map – for example, when the cartographer sends it back to the expedition headquarters – and makes the information it contains useful to other agents. For example, we can imagine the head of the expedition, back at the headquarters, receiving the map and studying its features. Noticing that the map indicates a river streaming through a narrow gorge, he might conclude that if the expedition had to cross that gorge, they would have to bring waterproof equipment.

We can generalise this insight about the cartographer, by noting that whenever we infer something about the world from some local observations, we must be able to place the observations within a system of coordinates that links them together. For example, if I toss a coin, observe that it lands Tails, and note down the outcome of the toss, I have made an observation which tells me something about what the world is like: namely, that it is a world in which the outcome of the coin toss is Tails.[11]

11 I will return to this point in Chapter 6, where I will give a more detailed discussion of the indexical element of empirical observations.

Similarly, Perry in *cartographer mode* uses the information he has collected about his own surroundings ('someone spilled sugar around the counter, I have been following this trail of sugar, but did not see whoever is producing it', etc.) to reconstruct the salient features of the world. Combined with his background knowledge of his own identity, this leads him to identify (w_3, a_P) as the actual centred world.

To sum up, the example of the cartographer tasked with producing a map of a territory illustrates how reasoning in the cartographer mode works. A subject reasoning in this mode infers some information about the world from the observation of her own circumstances. Moreover, she can use a system of coordinates to link successive observations (for example, spatial or temporal coordinates), pooling together the information from different observations.

3.2.2 *The pathfinder mode*

On the second reading of the messy shopper case, on the contrary, Perry is not seen as inferring something about the world from observations drawn from his surroundings. Instead, the inference goes in the opposite direction: from the information that he has about the world, he infers something about his own circumstances, which leads him to rule out some possible centres.

I will call the mode of reasoning that operates on this alternative reading of the messy shopper example the *pathfinder mode*. Again, an analogy may help to illustrate the operation of this mode. In this case, we can imagine a hiker who sets out on a walk, armed with a detailed map of the territory. Approaching a fork in the deep of a forest, his immediate surroundings do not tell the hiker where each path leads to. This information about the properties of his imme-

diate surroundings can, however, be inferred from the map, if the hiker is able to locate his own position with respect to it.

If the hiker has a reliable method to ascertain his own coordinates and identify them with points on the map, then the task of locating his own position with respect to the map will be an easy one. If, on the contrary, he is uncertain or has no reliable method to identify his coordinates, he may still try to reconstruct which is his position by comparing his present surroundings to what the map specifies for each point. If there is some unique feature that identifies his surroundings (for example, a landmark, or a particular sight, which is also marked on the map), then again he will be able to identify his position with respect to the map.

In this example, the map stands for the hiker's objective information about the world. When he sees the fork, he does not learn anything about the world that he did not already know from the map – the presence of a fork with those characteristics is recorded by the map. Rather, from comparing the map to his present circumstances, he can infer the details of his own current location and, in particular, its position relative to other objects.

We can generalise the analogy of the hiker, noting that when we have some prior knowledge about what the world is like, we can use it to locate ourselves with respect to other things in the world. In order to do this, we generally must rely on some identifying feature of our present circumstances. Imagine, for example, that Aron knows his friend Bella arrives exactly at 2pm. If Aron knows it is 2pm, he is able to infer that Bella arrives now; if, on the contrary, he sees Bella arriving he is able to infer that his current time is 2pm. In each case, either his background information about the time coordinate, or an identifying feature of his present circumstances, enables Aron to locate himself within the world.

The second intuitive interpretation of the messy shopper example corresponds to reasoning in pathfinder mode. In this case, Perry infers something about his own circumstances from what he knows about the world, together (presumably) with some identifying features of his own present circumstances. His immediate surroundings do not immediately reveal whether he is the messy shopper or not, but he knows that a particular agent has been walking around the counter, spilling sugar, following a path that coincides with his own. From this information about the world, Perry finally infers that he must be the messy shopper, and the actual centred world is (w_3, a_P).

3.2.3 *Learning and inferring from context*

To summarise, the main difference between the two modes of reasoning that are behind the two intuitive readings of the messy shopper case that I have identified is that on the *cartographer mode*, information about the properties of one's own circumstances is used to infer something about the world, whereas in the *pathfinder mode*, information about the world is used to infer something about one's own identity or location.

As the examples introduced in the last two sections illustrate, both modes of reasoning seem intuitively natural, in different circumstances. Sometimes, as in the case of the explorer tasked with producing a map of a territory, the cartographer mode describes the natural way in which the explorer would conduct her observations. Other times, as in the case of the hiker approaching a fork, the relevant information about the world is already known (as represented by the map), the pathfinder mode describes the natural way in which a subject would reason to identify the correct path to take. Both modes of reasoning therefore seem to occur quite naturally in our reasoning. This is perhaps

not surprising, given the fact that we are limited agents, whose interactions with our surroundings typically take place within a context, or are associated with a specific point of view, corresponding to our current position or identity. Thus our uncertainty can span two different dimensions: it can be about the world, reflecting our limited information about its objective features, or about ourselves, reflecting our limited information about the relative positioning of different locations, identities or perspectives on the world that we might occupy.

Sometimes, the two modes can also mix together, as illustrated by the following example. John has plans to go for dinner with Simon. The plan is that they will either meet at the local pub at 6:30pm, or they will meet at the Indian restaurant at 7pm. At some point in the late afternoon John receives a text from Simon that says 'meet me there in an hour!' What, if anything, can John infer from Simon's message? Let w_p be the possible world where John and Simon plan to meet at the pub at 6:30pm, and w_r the possible world where they plan to meet at the restaurant at 7pm. Suppose that John is not aware of what time it is at the moment he sees Simon's message. Since the plan is to meet either at 6:30pm or at 7pm, the message tells him that the present time is an hour before the planned meeting time, and so must be either 5:30pm (if they are meeting at the pub) or 6pm (if they are meeting at the restaurant). Taking these two times to be two possible centres at which John might currently be located, $(w_p, 5:30pm)$ and $(w_r, 6pm)$ are the two centred worlds that are compatible with John's current circumstances.

John's uncertainty in this example comprises both uncertainty about the world (he does not know whether the plan it to meet at the pub or at the restaurant) and about the centre (John does not know what time it is at the moment he receives the message). Moreover, in this specific example, it is clear that if John was able to independently learn either the current time or the plan, in other

words if he could access information about the centre or about the world, he would be able to infer which centred possibility corresponds to his actual circumstances. For example, if he could look at a clock and see that it is actually 5:30pm, the information about the centre that he would thereby acquire would enable him to infer that the plan is to meet at the pub. Conversely, if Simon sent an additional message to say 'oh, by the way: meet at the pub!', he would be able to infer that the current time is 5:30pm.

For another example to illustrate the mixing of the two modes of reasoning, imagine a cartographer who, during her explorations, loses track of her position. We can imagine that after being dropped on the hill top and conducting a series of measurements on the surrounding terrain, her instruments start to malfunction and as a result she can't say with any certainty how far her current position is with respect to the previous measurement. The cartographer now finds herself in a situation where both her current coordinates and the objective features of her environment are, to some extent, uncertain. As in Simon's dinner case, if she were able to learn either her current position, or some identifying features of the world, the cartographer would be able to infer the missing information (about the world, or about her current position, respectively) by applying either one of the two modes of reasoning.

As these two examples show, the two modes of reasoning are often interrelated in practical applications. There are, however, reasons to keep them distinct, because they capture two ways in which we can learn new information, or infer it from the context we are placed in. If we didn't have both modes of reasoning, it would be necessary to explain how these different modes essentially share a same underlying mechanism. Moreover, as the two examples highlight, both modes of reasoning correspond to the way we reason naturally when dealing with *de se* uncertainty. For this reason, I take it as a methodological desideratum that any semantic framework designed to explain the content of *de se*

expressions should be able to accommodate both modes of reasoning – or, if it fails to do so, it should give a good explanation for why one of the two modes of reasoning should be discarded.

3.3 THE SEMANTIC CONTENT OF DE SE EXPRESSIONS

In the previous sections, I have introduced – using Perry's example of the messy shopper – the notion of *de se* expressions and beliefs, and explained how they can figure in two different modes of reasoning, that I called the cartographer and the pathfinder modes. In this section, I turn to the question of what it is that we learn, when we learn some *de se* information. To answer this question, it is necessary to provide a semantic framework to analyse *de se* expressions and beliefs. I focus on two semantic frameworks that are prominent in the literature on *de se* beliefs, namely what I will call the Stalnakerian[12] and the Lewisian[13] accounts.

Since our purpose for giving a semantic framework here is to analyse *de se* expressions and beliefs, the exposition of each framework will be organised around three points:[14]

- What, according to the proposed framework, is the content or meaning of *de se* expressions;

12 See Stalnaker (1981, 2008, 2016).

13 See Lewis (1979) and Liao (2012). While both of these are closely inspired by the work of Robert Stalnaker and David Lewis, respectively, the present reconstruction does not aim at giving a faithful interpretation of the views expressed by these two philosophers.

14 This list does not aim to be exhaustive: there are certainly many other functions that a semantic framework for *de se* expressions and beliefs might perform, including explaining action, communication, and agreement. See Chapter 2, §2.3, and Ninan (2016) and Moss (2012), among others, for applications of semantic frameworks for *de se* expressions in these areas.

- How, on the proposed framework, we should represent the content of a subject's beliefs, including *de se* beliefs;

- In what ways the proposed framework explains how we reason about *de se* uncertainty. In particular, I will be interested in whether and how the proposed framework is able to accommodate the two modes of reasoning that I have identified in the previous section.

Moreover I will take the following to be two methodological desiderata. Firstly, a semantic framework should preferably give a unified account of the content of *de se* expressions and beliefs. Secondly, a semantic framework should account for both modes of reasoning about *de se* that were identified in the previous section, for the reasons expressed above.

3.3.1 *The Stalnakerian account*

Stalnaker argues that sentences containing indexical or demonstrative terms (or *de se* sentences) do not automatically correspond to fully-fledged propositions. Instead, for Stalnaker, a *de se* sentence may correspond to different propositions depending on the context at which it is uttered. As a consequence, the content of a *de se* sentence changes with respect to the context, which serves to fix the reference of indexical and demonstrative terms. Once the references are fixed, the resulting proposition corresponds to a set of (uncentred) possible worlds (see also Stalnaker (2014), ch. 5, and Moss (2012)). For instance, the content of the sentence 'I am the messy shopper', as believed by Perry at the time he stops to check his cart, corresponds on Stalnaker's account to the uncentred proposition containing all the possible worlds in which the individual designated as John Perry is the messy shopper. In our earlier formalisation, this corresponds to the unique possible world $\{w_3\}$. In this way, the

Stalnakerian account matches the traditional account of the semantic content of propositions, extending it to cover the content of *de se* propositions. On this account, the content of a belief is a set of uncentred possible worlds, or – in other words – an uncentred proposition.

The Stalnakerian account fits very naturally with the first reading of the messy shopper example, where the new information that Perry learns as he realises that he himself is the messy shopper is taken to be something about the world. What happens to Perry, according to this reading of the example, is that he finds out some uncentred information about the world, leading him to rule out some possibilities on the basis of the fact that their uncentred component does not match what he knows the actual world to be like.

On a traditional picture, the content of a subject's belief state is usually taken to correspond to the set of propositions that the subject believes. For example, Perry's belief state before he realises that he is the messy shopper, on the traditional picture, might contain all the propositions that Perry believes at the time, such as – for instance – that 'someone is making a mess', that 'the messy shopper has been walking around the counter', and 'there is sugar on the floor'. Stalnaker's position, however, is not completely reductionistic. On the Stalnakerian account, *de se* information has a role in explaining the nature of the relationship between a subject and the content of his or her beliefs. On this point, Stalnaker appears to agree with Perry and Lewis that *de se* information (essentially indexical information, for Perry) can not be deduced from a complete objective description of the world, and thus is not reducible to it.[15]

The Stalnakerian account uses centred worlds to model belief states. Formally, a belief state B is modeled as a set of centred worlds, where the w compo-

15 See e.g. (Stalnaker, 2008, p. 54): 'The role of the centers is to link the believer, and time of belief, to the possible worlds that are the way that the believer takes the world to be at that time, and to represent where, in those worlds, he takes himself to be.'

nent represents different ways in which the world might be, and the centred component c represents the identity and/or spatio-temporal location that the believer could, for all he or she knows, occupy.

What determines which proposition corresponds to the sentence 'I am the messy shopper' on the Stalnakerian account, as we have seen, is the context at which the sentence is uttered. Thus, on this account, the same sentence ('I am the messy shopper') corresponds to two different propositions when it is uttered by Perry and from that of Smith. Conversely, the same proposition (for example, that 'Perry is the messy shopper') can be believed by both agents, who could nonetheless express it using different sentences (for example 'I am the messy shopper' might be acceptable for Perry, while 'he is the messy shopper', uttered indicating Perry, might be acceptable for Smith).

Even though Smith and Perry believe the same proposition (namely, that Perry is the messy shopper, or w_3, it seems intuitively that they are nevertheless in two distinct belief states. On the Stalnakerian account, this intuition is cashed out in the following way. A belief state B is modeled as a set of centred worlds $\{(w_i, c_j)\}$ where each w_i specifies the objective features of the world, and c_j links this objective description to a specific perspective that is compatible with the current circumstances of the subject of the belief state B. In Stalnaker's words, this reflects the idea that 'the job of the center [is] to link the believer as he is in the world in which he has the beliefs to the person he takes himself to be in the world as he takes it to be' (Stalnaker, 2014, p. 115). So, on the Stalnakerian account, centres are not introduced to represent a dimension of a subject's uncertainty, but rather act as a link between the contents of a subject's beliefs (which, according to Stalnaker, are uncentred propositions), and their contextual circumstances.

I now turn to discuss a reading of the messy shopper example that appears to cause some problems for Stalnaker's account. As we saw earlier, the first reading of the example relies on the idea that Perry is never unsure about his own identity, so all the centred worlds compatible with Perry's circumstances are centred on the agent who is identified as Perry. However, as the second reading of the messy shopper example indicates, there can be cases where an agent's uncertainty is not so much about the world, but concerns their present location or identity – that is, there seem to be cases of genuinely *de se* uncertainty. What characterises this second type of uncertainty is that learning the missing *de se* information does not also automatically entail learning something about the world. In the second reading of the case of the messy shopper, for example, when Perry realises that he himself is the messy shopper he does not also contextually learn anything new about the world (since all the centred possibilities he entertained prior to the realisation coincided on w_3).

Since Stalnaker's account takes standard propositions (i.e. sets of uncentred possible worlds) to be the content of *de se* sentences, it *prima facie* faces a problem to account for this second type of uncertainty, which is not strictly speaking about the world, but about which of the possible centres matches a subject's current circumstances. On the Stalnakerian account, we can't explain this type of *de se* uncertainty by saying that the agent entertains two centred possibilities that coincide with respect to the possible world w, but differ with respect to the centre c, because these two possibilities correspond to the same proposition. In other words, an agent could not be uncertain about these two possibilities because, on the Stalnakerian account, they have exactly the same content. In light of the second reading of the messy shopper example, and the discussion under the pathfinder mode of reasoning, this is naturally a puzzling result for the Stalnakerian account. There seem to be cases where agents are uncertain about their own location or identity, but the Stalnakerian account seems to rule out this possibility.

Stalnaker proposes the following condition for belief sets, which he calls *propositionality*:

Definition 1 (Propositionality)**.** For any belief state B, and centred worlds (w, c), (w, c') that coincide with respect to their uncentred component w, $(w, c), (w, c') \in B$ if and only if $c = c'$.

In other words, propositionality says that any uncertainty about the centre is always also uncertainty about which uncentred possible world is actual. If propositionality holds, this implies that even when an agent has *de se* uncertainty, this uncertainty can ultimately be reduced to uncertainty about the world. Moreover, the way in which agents learn *de se* information is by acquiring new information about the world.

We can go back to the example of the hiker to illustrate how the Stalnakerian account works with *de se* uncertainty in this case. As in the previous example, we can imagine that Ted is hiking through a forest. He has a very detailed map of the area he is crossing, which he uses to guide his way. Arriving at a fork, he is unsure about whether he should keep to the left, continuing on the same trail, or whether he should turn right into a smaller path. He pulls out the map, but finds that there are two points that, for all he knows, are compatible with what he sees around himself.

A natural way to represent Ted's belief state would be to let w correspond to the objective information about the territory that is contained in the map, and let c and c' correspond to the two possible locations that are compatible with Ted's surroundings. The resulting belief set $B = \{(w, c), (w, c')\}$ violates propositionality, because c and c' are two distinct centres, but both are associated to the same possible world w within B. This way of representing Ted's belief state would therefore be rejected by the Stalnakerian account. Instead, a proponent

of this account should argue that if Ted is uncertain about these two locations, then c and c' must be associated with two distinct uncentred worlds, giving the revised belief set $B^* = \{(w, c), (w', c')\}$, where $w \neq w'$. In other words, if *de se* uncertainty is present, the map is not detailed enough to uniquely pin down the actual world w.

Stalnaker can justify this move by pointing out that if Ted is uncertain about which position he occupies within the world, then there is in fact something about the world which he does not know: namely, where the individual corresponding to himself is located. In this example, although we have assumed that the map is accurate, it does not contain all the information about the position of individual hikers walking through the territory. But this information should be part of a complete objective description of the world. Therefore, supposing for simplicity that Ted is the only hiker around, when he stops to consider his own position the two possible worlds that he considers differ both with respect to the location that he occupies within them (c, or c'), and with respect to whether the 'completed' map specifies the position of the hiker as being at c or at c'. Since there are two ways for the map to be completed, they correspond to two different uncentred worlds in Ted's belief set.

As this example shows, the Stalnakerian strategy to deal with cases of *de se* uncertainty involves turning *de se* statements (such as 'I am located at point c') into *de dicto* ones (which do not include indexical or demonstrative terms, such as 'the hiker is at point c'). This strategy entails that it is impossible for a subject to learn anything *de se* about his or her own location or identity, without at the same time also learning something about the world. In other words, new information is always *uncentred* information about the world.

Showing that the Stalnakerian account can be applied to the second reading of the messy shopper example is a bit more tricky. On that reading, Perry does

not learn anything new about the world, apart from the fact that he is the agent who is making a mess. The formalisation of this reading of the example that I gave in §3.1.2 is not immediately compatible with the Stalnakerian account, because it violates propositionality. So, we will need to find another way to capture the same intuition. Stalnaker could respond by saying that, on a fundamental level, Perry is not uncertain about his own identity, but he might still not know which properties apply to him in the actual world, including the names by which he happens to be known. For example, Perry might be for some reason oblivious to the fact that 'Perry' is a name that refers to himself, and so even if he believes that Perry is the messy shopper in the actual world, he does not know whether he himself is the messy shopper. For all that he knows, he wouldn't know any difference if he were either Jones, Smith or Perry. He does, however, know that *the very token thought that he is having* at the moment when he considers these possibilities is unique, in the sense that he is the only one who has it in the actual world. Call this thought token θ. Since he knows he has θ in the actual world, the question about which agent he is can now be reformulated in the following slightly different way, as which agent has the token thought θ in the actual world. There are three possibilities: either Jones has token thought θ, or Smith has it, or Perry has it. Therefore, we need to refine the description of the possible world w_3, which is now split into three further possibilities:

- w_3^J: Perry is the messy shopper and Jones has token thought θ;
- w_3^S: Perry is the messy shopper and Smith has token thought θ;
- w_3^P: Perry is the messy shopper and has token thought θ.

Given that Perry knows that he himself has the thought θ in the actual world, let 'ME' be the person who has token thought θ. His belief set will be equal to the set $B = \{(w_3^J, \text{ME}), (w_3^S, \text{ME}), (w_3^P, \text{ME})\}$. When he finally realises that he in

fact is the messy shopper, what Perry learns is effectively something about the world: namely, that w_3^J is the case, i.e. that it is the agent named Perry that has token thought θ.

To summarise, here are the basic ingredients of the Stalnakerian account:

- The contents of expressions or sentences (including *de se* expressions or sentences) are propositions, which on the Stalnakerian account correspond to sets of uncentred possible worlds;
- The content of a *de se* sentence or expression can vary with respect to the context at which it is uttered;
- A belief state is modeled as a set of centred worlds B, whose elements are all the centred worlds at which the agent, for all he or she knows, could be located;
- Belief sets satisfy the condition of Propositionality. This entails that uncertainty about where one is in the world is also always uncertainty about which world is actual. In other words, a belief set can never contain pairs of centred world that coincide on the w component, but differ with respect to c.

And a summary of the recipe:

> In order to reason about *de se* uncertainty, we do not need to posit the existence of other types of contents in addition to standard propositions. Contrary to standard propositions, that are true or false absolutely, *de se* sentences appear to be true in some contexts, but false at other contexts. This is because their propositional content varies, and is fixed by the context at which they are

evaluated. But once their content is fixed, they are equivalent to standard propositions, and can be treated in just the same way.

3.3.2 *The Lewisian account*

In contrast to the Stalnakerian account, the framework originally put forward by Lewis (1979) does not attempt to analyse the content of *de se* expressions by fixing their meaning relative to a context. Instead, Lewis argues that *de se* expressions, such as 'I am the messy shopper', or 'It is 4pm right now', correspond to what he calls *centred propositions*.

The issue with *de se* expressions, as we have seen, is that their truth value can change with respect to the context at which they are evaluated. For example, the truth value of 'It is 4pm right now' differs, depending on the actual time at which it is evaluated. On the Stalnakerian account, this is explained by the fact that the content of the *de se* expression varies with the context. On the Lewisian account, on the other hand, the content is kept fixed, and it is only the truth value which changes. It is a fairly standard view that the truth value of a proposition can vary across possible words – so we talk about a proposition being true at some worlds but not others. With Lewis, we have centred possible worlds, and the truth value of a proposition can vary across these. Of course, as time passes we are moving between centred possible worlds, and so the truth value of some propositions will change. More precisely, the idea which is at the basis of the Lewisian account is that the content of *de se* expressions is not essentially different from that of *de dicto* expressions. According to a standard picture, the proposition expressed by a *de dicto* sentence is just a function of the meaning of its constituent terms, and is fixed independently from a context of utterance. The truth value of a proposition, on the other hand, depends on

what the world is like – or, within a possible worlds framework, what possible world is the actual one at which the proposition is evaluated. For example, the sentence 'snow is white' corresponds, on a standard account, to the set of possible worlds $S = \{w : \text{snow is white at w}\}$. Conversely, the proposition 'snow is white' is true, when evaluated from the perspective of the actual world $w_@$, just in case $w_@$ is a member of S.

If we take a sentence like 'I am the messy shopper', however, this doesn't seem to express a proposition that automatically correspond to a set of possible worlds in the standard, uncentred sense. Nor does it seem possible to evaluate its truth value relative to an uncentred possible world: if a world w contains a messy shopper, it doesn't seem to follow that 'I am the messy shopper' is true at w. Lewis (1979) however argues that the content of a *de se* sentence can be expressed as a centred proposition, and the standard account of propositions can be extended to the centred case. According to Lewis, when Perry realises that he is the messy shopper, what he comes to believe is that he himself has a certain property. In other words, learning the *de se* proposition 'I am the messy shopper' means that Perry self-attributes the property of being the messy shopper.

Lewis introduces centred worlds to capture the content of such self-attributions of properties. On the Lewisian account, the *de se* sentence 'I am the messy shopper' (P) has a content that is independent from the context or the current circumstances of a subject that happens to utter it, and corresponds to the set of all the centred worlds (w_i, c_j) that are centred on an individual c_j who is the messy shopper in w_i. When Perry comes to believe that he is the messy shopper, therefore, he comes to believe that the centred world he currently inhabits is centred on an individual (himself) who is a messy shopper; in other words, that his actual centred world is a member of P.

The Lewisian account is able to accommodate naturally both modes of reasoning identified in the previous section. If we take Perry's background beliefs to include information about his own identity, as in the first intuitive reading of the messy shopper case presented in Section 2, learning the centred proposition 'I am the messy shopper' enables Perry to rule out from his belief set all the centred worlds that are not centred on a messy shopper (namely, (w_1, a_P) and (w_2, a_P)). This, in turn, enable him to single out (w_3, a_P) as the centred world corresponding to his circumstances.

If, on the other hand, Perry's background knowledge does not include an effective way to identify himself with a possible centre, but does include sufficient information about what the world is like, then learning the same centred proposition ('I am the messy shopper') enables him to rule out the centred worlds which are not centred on Perry himself (namely, (w_3, a_J) and (w_3, a_S)), again leaving the single actual centred world (w_3, a_P).

In the first case, we can see Perry as reasoning in the cartographer mode: learning something about his own circumstances leads him to update his beliefs about what the world is like. In the second case, on the other hand, we can see Perry as reasoning in the pathfinder mode: learning something the world leads him to update his beliefs regarding his own position, or who he takes himself to be within the world.

To summarise, the main ingredients of the Lewisian framework are the following:

- The contents of sentences and beliefs (including *de se* sentences and beliefs) are centred propositions, or sets of centred worlds;
- The belief state of an agent corresponds to the set of all the centred propositions believed by the agent;

And the recipe to reason about *de se* uncertainty is very simple: centred worlds should play the same role, within the Lewisian framework, as uncentred possible worlds play in the standard possible worlds semantics for propositions.

Based on my discussion, the Lewisian account of *de se* propositions seems to offer a much simpler explanation of the modes of reasoning involving *de se* uncertainty, while the Stalnakerian account appears to be incompatible with reasoning in the pathfinder mode. This, on a first blush, is a disadvantage for Stalnaker's account, as it does not satisfy the desiderata in §3.3. The Lewisian account, however, comes at the expense of modifying the standard account of the content of beliefs, admitting a refinement of possible worlds, in a way that has far reaching implications (as will be discussed in more detail in the following Chapters). [16]

3.4 CONCLUSION

In this chapter I have used a famous example, the case of the messy shopper first discussed by Perry (1979), to illustrate what *de se* uncertainty is. Building on two intuitively plausible readings of Perry's example, I identified two modes of reasoning that can come into play when a subject is uncertain about what the world is like (the *cartographer* mode), or when he or she is uncertain about the current perspective or location within the world (the *pathfinder* mode).

The two most successful accounts of *de se* sentences currently available, namely Stalnaker's and Lewis's account of centred contents, are not equally able to accommodate both modes of reasoning. While Lewis's account of centred propositions is compatible with both the cartographer and the pathfinder mode, Stal-

16 See Cappelen and Dever (2013) and Magidor (2015) for an extended analysis and critical discussion of the Lewisian framework on related points.

naker's account is only compatible with the cartographer mode, reducing all instances of self-locating uncertainty to uncertainty about the world.

To adjudicate between the two accounts, therefore, we need a further assessment. If Stalnaker's account is correct, and the condition that he calls Propositionality is true, then this would knock down the possibility of pathfinder mode.

In the next chapter, therefore, I turn to considering whether Stalnaker offers convincing reasons to accept propositionality, and whether his broader strategy to reduce the content of *de se* sentences to standard propositions is successful. In chapter 4 I will argue that Stalnaker's strategy fails, because it either makes it impossible for distinct agents to share a same context, or it is unable to explain how different agents who share the same context may assign a different truth value to the same *de se* expression. Therefore, we should look to the Lewisian account to provide a semantic framework for reasoning about *de se* uncertainty.

4

DE SE BELIEFS

In the previous chapters, I have discussed why we need *de se* beliefs and examined a particular way to model *de se* beliefs within a centred worlds framework. The discussion of Perry's messy shopper example highlighted how *de se* beliefs are essential to explain action. Moreover, as agents we seem to have a basic intuition that our experiences of the world are always given within a *perspective*. As agents, we situate ourselves within the world, and *de se* beliefs encode this individual perspective.

However, recognising the importance of *de se* beliefs raises a host of questions. In particular, it is unclear what exactly accounts for the special status of *de se* beliefs. If *de se* beliefs are special, it is also *prima facie* unclear how we should reason about *de se* information. Building on the discussion of the previous chapters, in this chapter I turn to consider more closely one particular answer to the problem of *de se* beliefs, according to which *de se* beliefs have similar propositional content as other types of beliefs, but have the special property of linking the agent with the content of his or her beliefs.

4.1 TWO QUESTIONS ABOUT DE SE BELIEFS

Perry's example of the messy shopper and other examples of cases involving self-identification or self-location are often used to illustrate how *de se* beliefs are special, and cannot be reduced to *de dicto* beliefs that are strictly about the world. As we saw through examples in the previous chapters, a belief is *de se* if it has the following characteristics: it is a belief about oneself, concerning one's identity or temporal or spatial location; it is typically expressed using an indexical sentence, such as 'I am the messy shopper' or 'The meeting starts now'; and finally it is not deducible from purely objective beliefs about the world. For example, Perry (1979) convincingly argues that 'I am the messy shopper' cannot be deduced from 'John Perry is the messy shopper', at least unless some other linking *de se* belief – such as 'I am John Perry' – is already in place.

Accepting that *de se* beliefs are in some way special and different from other types of beliefs, however, does leave open many further questions.[17] In this chapter I restrict my attention to two specific epistemological questions relating to the problem of *de se* beliefs, namely:

1. If *de se* beliefs are different from *de dicto* beliefs, what are the features that make them special?

2. Can there ever be genuinely *de se* uncertainty, or uncertainty that is purely concerning one's own identity or location?

17 For example, given the essential role they seem to play in explaining action (see e.g. Perry (1979), Ninan (2012)), one may ask whether non-human or group agents can have *de se* beliefs (List and Pettit, 2011). Moreover, *de se* beliefs play an essential role also in the way agents engage in communication, raising important questions in the philosophy of language (Stalnaker (2014), Ninan (2016), Stojanovic (2016)).

I am going to present and critically discuss a particular line of answer to these two questions. I will call this position Weak Acceptance, to distinguish it from two other positions which I will call Denial and Strong Acceptance.

4.1.1 *Three responses*

The problem of *de se* beliefs, as I have suggested in the previous section, has invited three main types of responses. The essential features of each response are summarised in table 1.

Proponents of Denial generally argue that the problem of *de se* beliefs is only a product of the possible worlds account of propositions (see Magidor, 2015; Cappelen and Dever, 2013). According to them, puzzling cases of *de se* beliefs can be adequately handled by any semantic framework that has the conceptual resources to solve puzzles about reference (more on this below). So, Denial would require a departure from the possible worlds framework that I have adopted and motivated in Chapter 2. This, however, comes at a significant cost: as will be discussed in Chapter 5, centred worlds provide a very natural framework to model uncertainty over de se possibilities, but the same is not true of other accounts of propositions.

The positions that I will call Strong and Weak Acceptance, on the other hand, are both compatible with a semantic framework that views propositions as sets of (centred or uncentred) possible worlds. The formulations of both positions that I will analyse in this chapter take *de se* beliefs to be represented as sets of centred worlds. The main difference between these two positions, however, is with respect to what they take the content of *de se* beliefs to be. For Strong Acceptance, the content of *de se* beliefs are simply centred propositions. So, on the Strong Acceptance account a belief and its content are formally the same

	Denial	**Strong Acceptance**	**Weak Acceptance**
Semantic framework	*Other*	*Possible (centred) worlds*	*Possible (centred) worlds*
De se beliefs modelled as	*Propositional attitudes*	*Sets of centred worlds*	*Sets of centred worlds satisfying Propositionality*
Content of *de se* beliefs	*Propositions (other)*	*Centred propositions*	*Uncentred propositions*

Table 1: Three positions on *de se* beliefs

thing, i.e. both correspond to a set of centred worlds. For Weak Acceptance, instead, the content of *de se* beliefs are uncentred (i.e. *de dicto*) propositions, and a belief and its content are formally distinct objects. On the one hand, a belief is modelled as a set of centred worlds that satisfies the condition of Propositionality (more on this below). On the other hand, the content of a belief corresponds to an uncentred proposition, i.e. to a set of ordinary (non-centred) possible worlds.

Now, let us consider more closely how each position can answer the two questions I have formulated above. According to Denial, the answer to the first question – about what if anything makes *de se* beliefs special – is that *nothing* makes them special. According to this position, *de se* beliefs, or beliefs concerning facts about oneself including one's identity and location, do not differ in any relevant respect from other types of beliefs that are simply about the world. Proponents of this position (including Magidor (2015) and Cappelen and Dever (2013)) generally argue that the problem cases, including the messy shopper and other examples of *de se* beliefs raised by Perry, should just be viewed as instances of a more general type of puzzles about reference, known

in the literature as 'Frege cases'. A famous example of such a puzzle is due to Frege (1892). Hesperus and Phosphorus are two names for the same object, a star that is visible in the sky both in the evening (hence the name Hesperus, or 'the evening star') and in the morning (hence its other name, Phosphorus, or 'the morning star'). Each name is associated with a different mode of presentation, and it is possible – indeed plausible – that someone observing Phosphorus in the morning sky might not know of the fact that it is the same star they observed in the evening, and which they then called Hesperus. Hence, the observer would be ignorant that the proposition 'Hesperus is Phosphorus' is true, even though it expresses a necessary truth because both names refer to the same object. Proponents of Denial argue that in cases involving *de se* beliefs, the indexical terms used to express these beliefs – such as 'I', 'now', or 'here' – refer to objects in the same way as a proper name such as 'Hesperus'. When I assert 'I am the messy shopper', I am asserting that 'I' and 'the messy shopper' refer to the same individual.

The answer to the second question – whether there can be genuine *de se* uncertainty – is less clear cut. On the one hand, since *de se* beliefs are not different from other types of beliefs, Denial has no difficulty in admitting that there can be uncertainty. On the other hand, the uncertainty involved is not specifically, or genuinely *de se*. For example, if I am uncertain about whether or not 'I am the messy shopper' is true, then, according to Denialists, this is simply because I lack the knowledge that 'I' (as uttered by me) and 'the messy shopper' in fact co-refer to the same individual.

Ultimately, I am not convinced that Denial is a satisfactory position with respect to the problem of *de se* beliefs. Since the primary focus of this chapter is the position of Weak Acceptance, I will not delve much deeper into a discussion of Denial, but I will just mention here a few main reasons why I do not think it can provide a satisfactory account of *de se* beliefs. First of all, in order

to be made workable, Denial must specify more precisely how indexical terms refer, and why the reference of these terms should be treated in a similar way as other referring expressions, such as proper names and definite descriptions. Without such an explanation, it is plausible to assume that indexical terms behave very differently from other referring expressions. Secondly, Denial does not explain the main issue raised by Perry, that is the irreducibility of *de se* beliefs. Analysing cases of *de se* uncertainty as instances of Frege cases hides the fact that *de se* beliefs appear to contain additional information, that is not possible to express using other means of presentation.

Moving on to the next position on *de se* beliefs, according to Strong Acceptance, the answer to the first question – what makes *de se* beliefs special – is that *de se* beliefs are different from other types of beliefs in virtue of their content. According to this position, ordinary, non-*de se* beliefs represent facts about the world and their contents are ordinary (*de dicto*) propositions. *De se* beliefs, however, represent a fine-graining of ordinary beliefs, and have as content *de se* propositions. In addition to accepting that *de se* beliefs are special in virtue of their content, Strong Acceptance also accepts the possibility of genuinely *de se* uncertainty. This type of uncertainty is different from ordinary uncertainty, because it concerns more finer grained possibilities than ordinary propositions. Strong Acceptance is my favoured position. It is also the position advocated by David Lewis[18], and the default underlying position in many accounts of puzzles of self-location.[19]

A significant alternative to Strong Acceptance is the position that I have called Weak Acceptance, which will be the main focus of the rest of this chapter. Weak Acceptance agrees with Strong Acceptance that *de se* beliefs are special, but disagrees on what features make them so. While according to Strong Acceptance

18 Lewis (1979, 1986, 2001)

19 See, e.g., Elga (2000) and Lewis (2001). See also Titelbaum (2016b) for an overview of the debate in this area.

de se beliefs are special in virtue of their content, according to Weak Acceptance the content of *de se* beliefs is no different from the content of any other types of beliefs, and is just ordinary propositions. I will say more on the details of this position in the following sections, but the main point of this position is that *de se* beliefs are different in the way that they represent this content. While *de dicto* beliefs represent propositions in an objective, third-personal way, *de se* beliefs can represent the same contents from a first-personal, subjective perspective. In other words, according to Weak Acceptance the contents of any type of beliefs are ordinary propositions, but the total cognitive state of an agent is not exhausted by the set of all the propositions that they believe – it also includes a 'perspective' from which these beliefs are held, represented by the agent's *de se* beliefs.

With respect to question 2, while Strong Acceptance agrees with the possibility of genuine *de se* uncertainty, Weak Acceptance denies this. According to this position, all uncertainty is ultimately about some content, and since the contents of *de se* beliefs are ordinary propositions, the uncertainty must correspondingly be just uncertainty about the ordinary propositions. In other words, Weak Acceptance maintains that that even when *de se* uncertainty seems to be present, it is always ultimately reducible to uncertainty about what the world is objectively like.

The aim of this chapter is to give a detailed analysis of Weak Acceptance, and ultimately argue that – despite its initial attractions – this is not a tenable position. For definiteness, my discussion will draw more specifically on the framework put forward by Robert Stalnaker, which I take to be the most compelling version of Weak Acceptance in the recent literature. The rest of the chapter is organised as follows. Section 4.2 outlines the answer that Weak Acceptance gives to the problem of *de se* beliefs and explains why it is a *prima facie* attractive position. Section 4.3 presents the two core assumptions that characterise

the position of Weak Acceptance, with a particular focus on the version proposed by Stalnaker. Section 4.4 presents the formal framework introduced by Stalnaker to represent *de se* beliefs and shows how the two core assumptions of Weak Acceptance can be articulated in this framework. Section 4.5 raises some objections to Stalnaker's framework. Section 4.6 considers whether these objections could be met by Weak Acceptance by changing the framework in appropriate ways, but shows that the problems with the formal framework ultimately derive from a deeper tension between the two core assumptions. Finally, Section 4.7 concludes that, based on the issues discussed in this chapter, Weak Acceptance is not an adequate account of *de se* beliefs.

4.2 REASONS FOR WEAK ACCEPTANCE

At least *prima facie*, Weak Acceptance has several attractive features. First of all, it accommodates the apparent irreducibility of *de se* beliefs, thus answering the issues about essential indexicality originally raised by Perry. Weak Acceptance maintains that *de se* beliefs are not reducible to beliefs about what the world is objectively like, because they also encode some further facts concerning the particular perspective on the world that agents occupy. These further perspectival facts function as a link between a subject and the content of his or her own beliefs, and are essential to explaining reasoning and action. Weak Acceptance does not deny the presence of *de se* uncertainty, but argues that all its instances are ultimately reducible to ordinary objective uncertainty. Apparent *de se* uncertainty intuitively arises in all the cases in which the perceived features of an agent's own perspective on the world are compatible with different spatial locations, times, or identities that the agent – for all he or she knows – might occupy. As will be discussed in more detail below, this feature allows Weak Acceptance to give an account of cases in which agents intuitively

have some *de se* uncertainty, such as the ones illustrated by the examples in the previous chapters.

Unlike a proponent of Denial, a proponent of Weak Acceptance is not committed to denying the intuitively plausible view that the particular perspective that each agent occupies within the world is relevant to the way agents act and reason based on their beliefs. In the case of the messy shopper, for example, Perry's realisation that *he himself* is the messy shopper is what induces him to pause and check his cart. Similarly, Weak Acceptance seems to vindicate the basic intuition that each individual experiences the world from their unique individual perspective.

Another feature that makes Weak Acceptance an attractive answer to the problem of *de se* beliefs is that it is compatible with standard accounts of propositions and updating. Unlike Strong Acceptance, Weak Acceptance is compatible with the view that the contents of beliefs correspond to ordinary propositions, understood as sets of possible worlds. This can be regarded as an advantage, as it means that Weak Acceptance is still compatible with standard accounts of belief change, in particular it is compatible with a Bayesian account of updating. By contrast, Strong Acceptance is usually taken to be incompatible with standard versions of Bayesian conditioning (see Chapters 5 and 6), and therefore proponents of Strong Acceptance may need to provide alternative accounts of belief updating (see Titelbaum (2008, 2016b) for a discussion of recent attempts). Contrary to this general opinion, I think that Strong Acceptance is consistent with standard norms of Bayesian reasoning, and in Chapter 6 I will present my own solution to the problem.

4.3 TWO CORE ASSUMPTIONS

In this section, I present the two core assumptions that characterise the position of Weak Acceptance. I will focus in particular on the work of Stalnaker. The two core assumptions that I identify, namely the Non-deducibility of *de se* beliefs and the condition of Propositionality that is assumed to provide a link between *de se* beliefs and propositional contents, spell out more clearly the details of the answers given by Weak Acceptance to the two questions about *de se* beliefs that I raised in Section 4.1.

4.3.1 *The Non-deducibility of de se beliefs*

Weak Acceptance agrees that there is something special about *de se* beliefs. Stalnaker agrees with Perry that *de se* beliefs are not reducible to beliefs about the world. For Perry, as we have seen, *de se* beliefs are needed to fill the motivation gap. For Stalnaker, the role of *de se* beliefs is to link a subject with the content of his or her own beliefs. On Stalnaker's account, as we will see in a moment, an agent's cognitive state is not fully exhausted by the objective facts that he or she believes. An additional feature of belief states is that they represent propositional contents from a specific perspective, that is the one occupied by the agent who has the beliefs. In this framework, *de se* beliefs are what is used to represent ordinary propositional contents from a specific perspective.

De se beliefs, according to Stalnaker, have the role of linking a subject to the content of his or her own beliefs. To illustrate this point, Stalnaker (2016) uses the following example involving two people, a shared context and no *de se* uncertainty.

> Albert is in the kitchen and Boris is in the basement. Each knows who and where he is, and who and where the other is, so there is no self-locating ignorance. They each know all the same objective facts about their respective locations in the house, but there is still a difference in their epistemic states, a difference in their perspectives on the world.

The example of Boris and Albert serves to illustrate an intuitively common predicament: several people can share a common context, have all the same beliefs about the circumstances surrounding them, and yet all of them have a different perspective on the world. Although Albert and Boris have access to all the same information about the house and their respective locations, each of them has a different perspective on the same facts. To further drive this point home, Stalnaker continues:

> Suppose [that] a representation [of the contents of Boris and Albert's common state of belief] contained all the information about the beliefs of any person who is in the cognitive state that Boris and Albert are both in. Let x be any person in that state. Where does x believe himself or herself to be? It is clear enough from the description of the scenario that Boris believes he is in the basement and Albert believes he is in the kitchen, but these are further facts that are not reflected in the common set of propositions that is what each of them believes.

The example with Boris and Albert is supposed to illustrate how *de se* information is essentially different from objective information about what the world is like. It is possible to have complete information about the world, including the cognitive states of all the agents that are in it, without being able to deduce from this any *de se* information about who or where one is. In other words, *de*

se beliefs are not deducible from objective beliefs. For this reason, Stalnaker accepts the following assumption:

Assumption 1 (Non-Deducibility)**.** *De se* information cannot be deduced from a purely objective representation of the world. In particular, *de se* beliefs cannot be deduced from a set of objective propositions (in the absence of any further premises).

4.3.2 *Propositionality*

Weak Acceptance need not rule out that agents can, sometimes, be uncertain about some *de se* information. This would be *prima facie* implausible, as agents are often uncertain about where, when or – sometimes – who they are. For example, in the case of the messy shopper, Perry is initially uncertain about whether *he himself* is the messy shopper.

So, if the content of *de se* beliefs can be expressed using a first-personal indexical sentence by the agent who holds that belief, it is plausible that agents are sometimes uncertain about whether the sentences 'I am happy' and 'I am the messy shopper' are true. Weak acceptance does not have to deny that uncertainty is present in these cases. Instead, what it argues is that the relevant uncertainty is about what the world is like, and does not genuinely involve *de se* beliefs. According to Stalnaker, although sentences like 'I am happy' and 'I am the messy shopper' can be used to express *de se* beliefs, their content is in any case a proposition corresponding to a particular way in which the world might objectively be. Which particular proposition is the content of a specific utterance of 'I am the messy shopper' depends on the context. If it is Perry that utters it – and here let us assume (for now) that Perry knows that 'John Perry'

designates himself, then the content of 'I am the messy shopper' corresponds to the set of all the possible worlds in which John Perry is the messy shopper.

When Perry is uncertain about whether he is the messy shopper, this is because there are two ways that the world could be that he takes to be possible. In one possibility, the individual who is the messy shopper coincides with the individual that Perry take himself to be; while in the other possibilities, a different individual is the messy shopper. So, while Perry is uncertain about whether *he himself* actually is the messy shopper, this is not because he does not know who he is – he does not doubt at any point his own identity. Rather, the uncertainty comes from the fact that he lacks a crucial piece of information about the world, namely that the individual who is named John Perry is the messy shopper. If he had that additional information, given that he already knows that he is John Perry, it would also resolve his uncertainty about his *de se* beliefs.

This idea about *de se* uncertainty always being reducible to uncertainty about the world is captured by Stalnaker with the following condition, that he calls Propositionality:

Assumption 2 (Propositionality)**.** Uncertainty about *de se* information is always also uncertainty about the world.

In line with Weak Acceptance, Stalnaker's Propositionality condition ensures that the content of an agent's *de se* beliefs can always be mapped to standard propositions, expressing what the world is objectively like. Propositionality applies to *de se* uncertainty about time, and this allows Stalnaker to retain a fa-

miliar account of belief updating over time for agents who are uncertain about their *de se* beliefs, as the following example illustrates:

Example 1. Waking up in the middle of the night, Emily is unsure about what time it is. She remembers that her neighbour said he was leaving for a flight very early in the morning, but she doesn't recall if he told her that he would leave at 4am or 5am. She listens for a moment and notices that some birds are already singing. From this, she concludes that it must indeed already be 5am.

In this example, Emily has some *de se* uncertainty, as she does not know when she wakes up whether it is 4am or 5am. She believes that the noise which caused her to wake up was made by the neighbour as he was leaving, but since she forgot the exact time he told her, both possibilities are compatible with her present circumstances, as far as she knows now. So, both 'It's now 4am' and 'It's now 5am' represent live *de se* possibilities for Emily as she wakes up. Both also correspond to two different possible worlds.

According to Weak Acceptance, Emily's uncertainty in this example is *de se* in the sense that 'It's now 4am' and 'It's now 5am' correspond to two different time points at which Emily may locate herself in the world. However, this uncertainty is not *genuinely, or irreducibly de se*, because the two possibilities also correspond to two different ways that world might be objectively like. In one possibility, which corresponds to 'It's now 4am', Emily's neighbour leaves the house making some noise at 4am in the morning. In the other possibility, corresponding to 'It's now 5am', the neighbour leaves at 5am. These two possibilities differ not just in *de se* respects, about where Emily takes herself to be located at present (4am or 5am, respectively), they also correspond to two different ways that the world might be. If she could remember that the neighbour told her that he would in fact leave at 5am, Emily would no longer be uncertain about her *de se* beliefs. Therefore, Emily's uncertainty, although

it involves some *de se* beliefs, is in fact reducible to uncertainty that is strictly speaking about the world. As she wakes up, therefore, her beliefs are compatible with two different possible worlds, one where the neighbour's departure (and her own awakening) take place at 4am, and a different one in which the same events take place an hour later. Hearing the birds singing gives Emily an additional piece of evidence. Since she believes that these birds do not normally sing too long before dawn, she concludes by conditionalising on this new piece of evidence that the second possibility must be correct, and it must now be 5am.

4.4 STALNAKER'S FRAMEWORK

As seen in the last two sections, Non-Deducibility and Propositionality are the two core assumptions of Stalnaker's version of Weak Acceptance. Non-Deducibility affirms that *de se* beliefs are not reducible to beliefs about what the world is objectively like. Propositionality, on the other hand, is a condition meant to establish a correspondence between an agent's *de se* beliefs and the objective content of those beliefs.

In this section I present the formal framework that Stalnaker introduces to represent the belief states of agents. The presentation in this section is based on the Appendix to ch. 3 of Stalnaker (2008), where Stalnaker sketches a formal statement of the semantic framework he proposes. This will provide a useful background to articulate more precisely the essential features of Stalnaker's proposal.

Stalnaker takes as a starting point a standard possible world framework to represent the content of propositions. In this framework, propositions are taken to be sets of possible worlds. For example, the proposition P: 'Water freezes

at 0^oC' corresponds to the set of all possible worlds in which P is true. To this standard account of propositions, Stalnaker adds a further element to represent *de se* beliefs. Following Lewis (1979), he uses *centred worlds* to model *de se* beliefs.

Let W be a set of possible worlds and C be a set of possible centres. The elements of C can be specified in various ways, but should generally be taken to be n-tuples of an agent, a time, and a spatial location, identifying a particular perspective. *Centred worlds* are simply ordered couples (w, c) of a possible world $w \in W$ and a centre $c \in C$. Let Ω be the set of all the centred worlds (w_i, c_j) that are *logically possible*, that is that satisfy the minimal condition that the centre c_j exists within w_i.

The centred worlds in Ω represent all the possible predicaments in which an agent could be. For example, if w_p is a possible world where proposition P is true, and c_x is a centre within w_p, then (w_p, c_x) represents the predicament of an agent within w_p, who is located at c_x.

Wherever they might be located, agents have beliefs about the world that typically depend in some measure on the evidence that is available to them. This is captured in Stalnaker's framework by an epistemic accessibility relation R, defined over the centred worlds in Ω. For example, an agent located at (w, c) might consider his or her own evidence to be compatible with some other possibilities, say (w', c') and (w', c'), which are said to be *R-accessible* from (w, c). The epistemic accessibility relation R has the following properties:

Property 1 (Transitivity)**.** For any three centred worlds (w, c), (w', c'), (w'', c'') in Ω, $(w, c)R(w', c')$ and $(w', c')R(w'', c'')$ imply $(w, c)R(w'', c'')$.

Property 2 (Seriality)**.** For any $(w, c) \in \Omega$, there is some $(w', c') \in \Omega$ such that $(w, c)R(w', c')$.

In other words, seriality means that every centred world $(w, c) \in \Omega$ is R-related to something else. There is no logically possible centred world which is epistemically cut off from everything. In other words, there is no logically possible centred world (w, c) that is a 'blind spot', in the sense that an agent located at (w, c) would not consider any centred world (including also (w, c)) as possible.

Property 3 (Left-Euclidean)**.** For any three centred worlds (w, c), (w', c'), (w'', c'') in Ω, $(w, c)R(w', c')$ and $(w, c)R(w'', c'')$ imply $(w', c')R(w'', c'')$.

The property of being left-Euclidean means that if two different centred worlds are epistemically accessible from a third centred world, then they must also be mutually epistemically accessible between themselves. For example, suppose that it is actually 4:30am and Emily is awake, but she believes that it must be either 4am or 5am. The centred world corresponding to Emily's actual circumstances is $(w, c_{e,4:30am})$. The epistemic alternatives accessible from this centred world are $(w, c_{(\text{e, 4am})})$ and $(w, c_{(\text{e, 5am})})$. Since $(w, c_{(\text{e, 4:30am})})R(w, c_{(\text{e, 4am})})$ and $(w, c_{(\text{e, 4:30am})})R(w, c_{(\text{e, 5am})})$ both hold, the left-Euclidean property requires that $(w, c_{(\text{e, 4am})})R(w, c_{(\text{e, 5am})})$ also holds.

The three above properties imposed on the epistemic accessibility relation R induce the modal logic known as D45. To these, Stalnaker adds a further condition to capture the requirement of Propositionality:

Definition 2 (R-Propositionality)**.** For any three centred worlds (w, c), (w', c'), (w'', c'') in Ω, if *both* $(w, c)R(w', c')$ and $(w, c)R(w'', c'')$ *and* $w' = w''$, then $c' = c''$.

We have seen previously that Propositionality asserts that uncertainty about the centre is always also uncertainty about which possible world is actual. R-propositionality translates this condition at the level of the epistemic accessi-

bility relation R, constraining it in the following way: any two distinct centred worlds that are R-related between themselves must differ with respect to their uncentred component w. In other words, definition 2 requires that any two centred worlds y and z that are epistemically accessible from some other centred world x can only differ with respect to their centre if they *also* differ with respect to their possible world component. This ensures that in no case the epistemic possibilities accessible from a centred world x differ only with respect to the centre. Each epistemic possibility that is accessible from x must be characterised by a different possible world component.

Finally, building on the relation R Stalnaker introduces the concept of a *belief state*, denoted Bel:

Definition 3 (Belief state)**.** Given a centred world $(w, c) \in \Omega$, $Bel_{(w,c)}$ is the set of all the centred worlds $(w', c') \in \Omega$ such that $(w, c)R(w', c')$.

In other words, if (w, c) is the centred world corresponding to an agent a's current circumstances, then $Bel_{(w,c)}$ is the set of all the centred worlds that are epistemically accessible from (w, c), and which represent all the possibilities that are compatible with a's present evidence.

I will conclude this section with a few remarks on the framework presented so far, before we move on. Within Stalnaker's framework, the belief state of an agent depends both on the centred world at which the agent is currently located (which is called the *base centred world*) and on the epistemic accessibility relation R, which determines the set of centred worlds that the agent considers to be 'live possibilities' given his or her evidence. R is defined as an exogenous feature in this framework and is not agent-specific. It should be interpreted as specifying which centred worlds are compatible with the evidence that is available to an agent occupying any given centred world.

While belief states in this framework are represented by sets of centred possible worlds, the condition of Propositionality ensures that they can always be mapped to ordinary propositions. The *content* of a belief state Bel is defined as the ordinary proposition to which Bel is mapped. As evidenced by the example of Boris and Albert, two different belief states Bel and Bel' may be mapped to the same set of ordinary possible worlds, i.e. to the same ordinary proposition. In such cases, Bel and Bel' have the same content, but differ in the perspective from which they represent it.

A consequence of Stalnaker's framework is that any two belief sets Bel and Bel' are either identical or disjoint. More formally, Stalnaker's framework implies the following:

Proposition 1. For all (w,c) and $(w',c') \in \Omega$, either $Bel_{(w,c)} \cap Bel_{(w',c')} = \emptyset$, or $Bel_{(w,c)} = Bel_{(w',c')}$.

In other words, given any two centred worlds (w,c) and (w',c'), the intersection of the belief sets generated by each is either empty or coincides with the union of both belief sets (that is, (w,c) and (w',c') generate the same belief set). What is ruled out is that two belief sets with non-empty intersection do not coincide.

Proof. Suppose that $Bel_{(w,c)} \cap Bel_{(w',c')}$ is non-empty. Then there exists (w^*,c^*) such that $(w,c)R(w^*,c^*)$ and $(w',c')R(w^*,c^*)$. Now take any $(w^\dagger,c^\dagger)$ such that $(w,c)R(w^\dagger,c^\dagger)$. We need to show that we also have $(w',c')R(w^\dagger,c^\dagger)$. Since $(w,c)R(w^*,c^*)$ and $(w,c)R(w^\dagger,c^\dagger)$, the left-Euclidean property implies that $(w^*,c^*)R(w^\dagger,c^\dagger)$. Since $(w',c')R(w^*,c^*)$, Transitivity implies that $(w',c')R(w^\dagger,c^\dagger)$, as required. For symmetry reasons, the argument that $(w',c')R(w^\dagger,c^\dagger)$ implies $(w,c)R(w^\dagger,c^\dagger)$ is perfectly analogous, completing the proof. □

4.5 SOME PROBLEMS FOR STALNAKER'S FRAMEWORK

In this section I discuss some problems that arise when we use Stalnaker's framework to model de se beliefs.

4.5.1 *Intra-world ignorance*

One of the standard objections moved towards Weak Acceptance focuses on the condition of Propositionality.[20] Weber (2015) argues that cases where Propositionality is violated are possible and, in fact, quite common. He describes several counterexamples to Propositionality, calling them cases of Intra-World Ignorance (IWI for short). All cases of IWI share two characteristic features. Firstly, the agent has some *de se* uncertainty about his or her own current circumstances. In the previous example, for instance, Emily is woken up by some noise, which she was expecting, but is presently unsure about whether it is 4am or 5am. This uncertainty corresponds to two different time locations that she thinks would be compatible with her current circumstances. Secondly, in all cases of IWI, there is some pair of centred worlds (w^*, c) and (w^*, c') which the agent considers compatible with her own current circumstances, and which coincide on w but not on the centre that the agent occupies in each possibility. For example, in Emily's case, w^* could be a possible world in which Emily is woken up twice during the night by some similar noises, and in each case when she awakes she is uncertain about the time and does not remember any past awakenings. Since each of the two awakenings in this scenario would be indistinguishable to Emily, it seems that if her belief set contains one, it should

20 See Ninan (2012, 2016), Weber (2015).

also contain the other possibility. This, however, would be a straightforward violation of Propositionality.

Note that for Weber's objection to Propositionality to go through, it is not required that the pair of indistinguishable centres satisfying IWI are located in the actual world. All that is necessary is that there are *some* possibilities within the agent's belief set that satisfy IWI, even if these possibilities are not, in fact, actual. Excluding this possibility might be difficult within Stalnaker's framework, as the discussion under 4.5.3 below will suggest.

Responding to Weber's objection, Stalnaker (2016) argues that Propositionality should be taken as a starting assumption for his position, rather than as a proposition in need of being proved. According to Stalnaker, Propositionality is intuitively plausible and, in addition, is consistent with his own other views on metaphysics and the theory of reference. I will return on some of the deeper issues underpinning Propositionality in section 4.6, but for the moment entering a detailed discussion of the latter would take us well beyond the scope of this chapter. It is nevertheless worth noting that Stalnaker's reply does little to directly answer Weber's point.

4.5.2 *Radically mistaken de se beliefs*

Another difficulty for Stalnaker's framework that I wish to raise relates more closely to the nature of the epistemic accessibility relation R. The way in which Stalnaker defines it, R is not a reflexive relation, that is it does not generally satisfy the following property:

Property 4 (Reflexivity)**.** For any $(w,c) \in \Omega$, $(w,c)R(w,c)$.

If reflexivity holds, any logically possible centred world is epistemically accessible to itself. This in turn entails that for any (w, c), the belief set $Bel_{(w,c)}$ always contains (w, c). In other words, reflexivity entails that an agent a could not be fundamentally mistaken in their own beliefs, in the following sense: it could not be the case that the actual world and the centre that corresponds to a's current circumstances are not included in a's belief set.

Conversely, non-reflexivity entails that an agent's belief states could contain what I will call some radical mistakes. Assuming that (w, c) corresponds to the agent's current circumstances, these radical mistakes can be of two sorts. An agent's belief set could not contain any possibility corresponding to the actual world w. This would be the case if all the epistemically accessible centred worlds from (w, c) do not contain w. In this case, the agent's beliefs would be mistaken in the sense that he or she would believe that the actual world was impossible (that is, incompatible with her evidence). To give an example, we can imagine an agent who, for some reason, believes that the Sun actually orbits the Earth. None of the centred worlds in his or her belief set are compatible with the proposition 'The Earth orbits the Sun'. Since the latter is a true proposition, none of the possibilities in the agent's belief set contains the actual world.

However, there is also a second sense in which an agent's beliefs can be mistaken, if R is not required to be reflexive. An agent's belief set could contain the actual world w, but the agent might fail to locate themselves accurately within w. To illustrate this possibility, we could again use Perry's messy shopper example. Suppose that Perry correctly believes that two possible worlds, w and w', are compatible with his current circumstances. Furthermore, within the possible world w the individual corresponding to John Perry is the messy shopper, while within a different possible world w' some other individual, say Jones, is the messy shopper. As a matter of fact, w is the actual world, and Perry

is indeed occupying the centre identified by the individual called John Perry, denoted c_p, but – here is the catch – he is mistaken in his own identification. We can represent this by saying that Perry's belief set at (w, c_p) is $Bel_{(w,c_p)} = \{(w, c_x), (w', c_x)\}$, where $c_x \neq c_p$. This belief set is rationalised by an epistemic accessibility relation that specifies that (w, c_p), (w, c_x) and (w', c_x) stand in the following relations between themselves: $(w, c_p)R(w, c_x)$, $(w, c_p)R(w', c_x)$, $(w, c_x)R(w', c_x)$ and $(w', c_x)R(w, c_x)$. Under this specification, R satisfies transitivity, seriality (each of the three centred worlds concerned is R-related to at least one centred world) and the left-Euclidean property. Moreover, this specification of R satisfies Propositionality, since the centred worlds that are contained in Bel all differ with respect to their possible world component. This kind of belief set is therefore supported by Stalnaker's framework. Nevertheless, this seems problematic for at least two reasons.

To see the problem, first consider that Perry seems to believe something that is false – namely, that if w is the actual world, then he is not the messy shopper. This would suggest that the content of Perry's beliefs is some false proposition. However, the problem is that the Weak Acceptance view is unable to account for what is wrong in Perry's beliefs, because it cannot pinpoint what proposition he is ignorant of. On the Weak Acceptance view, it is not the case that Perry believes a false proposition, since his belief set contains the actual world. And Non-deducibility rules out that the objective features of the world fix which are the right *de se* beliefs, since these are considered to be some further facts independent as illustrated by the example of Boris and Albert.

How could Stalnaker respond? One way to handle this problem would be to impose reflexivity on the epistemic accessibility relation R. This, however, would be undesirable, because it would also preclude us from modelling the beliefs of agents that are simply mistaken about some objective features of the world, such as the beliefs of an agent who mistakenly believes that the proposition

expressed by 'the Earth does not orbit the Sun' is true. Since this proposition is false in the actual world, requiring the reflexivity of the epistemic accessibility relation means that at least one possibility that is consistent with 'the Earth does not orbit the Sun' being false must be contained in the agent's belief set, thus effectively ruling out that the agent could fully believe a false proposition. In other words, requiring the reflexivity of the epistemic accessibility relation would only solve the problem of mistaken *de se* beliefs by removing altogether the possibility of having mistaken beliefs.

Another way to solve the problem of mistaken *de se* beliefs, while getting around the issues posed by reflexivity, would be to modify the condition of Propositionality. A possible amendment of this condition, which would exclude the problematic cases discussed in this section, would be the following:

Definition 4 (R-Propositionality (2))**.** For any two centred worlds $(w,c),(w,c') \in \Omega$, if $(w,c)R(w,c')$ then $c = c'$.

Definition 4 is a considerable strengthening of definition 2, as definition 4 requires that all the centred worlds that are in $Bel_{(w,c)}$ differ among themselves with respect to their possible world component. Assuming that (w,c) is the centred world at which the agent is currently located, 4 also entails that one of two conditions holds: either a) $(w,c) \in Bel_{(w,c)}$, so the actual centred world is a 'live possibility' and is the only one contained in the agent's belief state that contains the actual world w; or b) if $(w,c) \notin Bel_{(w,c)}$, then every centred world (w',c') that is contained in $Bel_{(w,c)}$ does not have the actual world w as a component. In other words, 4 ensures that the centred world at which the agent is currently located is the only one that corresponds to the actual world w within the agent's current belief set.

Like the requirement of reflexivity, this strengthened version of Propositionality seems unwarranted for two main reasons. Firstly, it is hard not to see it as an *ad hoc* adjustment, as the only motivation for requiring this stronger condition would seem to be in order to avoid the problem of mistaken *de se* beliefs.

Even if this were not the case, though, the proposed modification of Propositionality would still be problematic, because it would entail that an agent who has mistaken *de se* beliefs must always be mistaken about the objective features of the world. In other words, it means that it is impossible for an agent to mis-locate themselves within the actual world. Again, it is unclear why this should be required, since by Non-deducibility it is assumed that *de se* beliefs are further facts that are independent from an objective representation of what the world is like.

4.5.3 *Overlapping belief sets*

Another problem that I wish to raise for Stalnaker's framework concerns the way in which it deals with the interaction between objective and *de se* uncertainty. The worry here is that Propositionality, together with the way in which the epistemic accessibility relation R is constructed, rule out the possibility that the belief sets of two distinct individuals might contain a common element. If this happens, then by Proposition 1, the two belief sets would have to be identical. However, cases where two agents have belief sets that overlap – but do not coincide – are both plausible and widespread. Moreover, Stalnaker's own framework and his analysis of examples of *de se* uncertainty seems to easily allow the construction of such cases.

To illustrate this problem, I will use a modification of an example initially proposed by Lewis (1979) and reprised several times by Stalnaker.[21] Imagine that there are two gods, inhabiting the same world, but one of the gods is located on the highest mountain, while the other god is located on the coldest mountain. Lewis (1979) originally argued that the two gods could both be omniscient with respect to all the objective features of the world that they inhabit (thus narrowing down the possibilities to the actual world $w_@$) and at the same time be ignorant about some *de se* information, so that despite their objective *de dicto* omniscience, each god could be ignorant of their own location within the world.

Lewis's example of the two gods squarely contradicts Propositionality, and Stalnaker's reply has been to argue that the two gods cannot, in fact, be omniscient if they lack some *de se* information. On Stalnaker's analysis of the example, each god is uncertain about which of two possible worlds is actual: one where he himself is on the coldest mountain, or one where he himself is on the highest mountain. Contrary to what Lewis argued, according to Stalnaker these must be understood as two distinct *de dicto* possibilities.

To support this conclusion, Stalnaker argues that the gods could identify the god on each mountain by giving him a proper name – e.g., naming 'Castor' the god on the highest mountain, and 'Pollux' the god on the coldest mountain. Suppose now that you are one of the gods, and you are uncertain about whether you are on the highest mountain or on the coldest mountain. If you were to learn that you are Castor, this would imply that you are on the highest mountain. So, on Stalnaker's account your belief set can be represented as $Bel_{(w_1,c)} = \{(w_1, c), (w_2, p)\}$, where w_1 is the possible world where you are Castor and Castor is located on the highest mountain, and w_2 is the possible world where you are Pollux and Castor is located on the highest mountain. w_1 and

21 See Stalnaker (2008, 2014, 2016)

w_2 count for Stalnaker as two distinct *de dicto* possibilities, because he thinks 'I am Castor' (or 'I am Pollux') also have *de dicto* content: they pick out the possible worlds where Castor (respectively, Pollux) is the god who has the same token thoughts as you do at the time you utter the sentence.[22]

Your counterpart god, in both Lewis's and Stalnaker's versions of the example, suffers from exactly the same uncertainty as you. What will his belief set be like? Intuitively, the content of his belief set should be the same as the content of your belief set, since both of you – by assumption – share the same propositional knowledge about the world[23]. As we assumed that you are Castor in the actual world w_1, this means that the other god must be Pollux and his belief set will be $Bel_{(w_1,p)} = \{(w_1, p), (w_2, c)\}$. Note that the belief sets of the two gods, as described, are mapped to the same ordinary proposition (the union of w_1 and w_2), but are disjoint. Both you and your godly counterpart are ignorant of who you are, and also ignorant of which possible world is actual. This way of analysing the example of the two gods satisfies Propositionality, since the *de se* uncertainty about each god's identity and location is mapped to ordinary uncertainty between two possible world alternatives.

Stalnaker's analysis of the two gods example, however, turns on the possibility of performing two related things: first, it must be possible for the gods to introduce proper names to designate objects in the world (including themselves). Second, the assigned names must be part of the objective features of the world, so that 'Castor is the god on the highest mountain' is an ordinary proposition that is true at the actual world w_1. However, these two things can come apart. To see this, again suppose that you are one of the two gods, and you are uncertain about whether you are the god on the highest mountain or the one on the

22 I will come back to this point later on in the discussion under §4.6.

23 Similarly to the Boris and Albert example, here we are assuming that the belief states of both gods have the same ordinary propositional content, although they might differ with respect to the *de se* information that they encode.

coldest mountain. You name the god on the highest mountain 'Castor', and the god on the coldest mountain 'Pollux'. Your belief set, as before, can be represented as $Bel_{(w_1,c)} = \{(w_1, c), (w_2, p)\}$. In other words, as before, there is some *de se* information that you lack (who you are and where you are located) that can be mapped to your uncertainty about which of w_1, w_2 is the actual world.

But suppose now that your counterpart god takes a slightly different approach to tackle his own uncertainty. He names *himself* 'Pollux', so the piece of *de se* information that concerns his own identity is fixed – he has no uncertainty about his own identity, he just is 'Pollux'. He is still ignorant about two objective possibilities, namely whether Pollux is on the coldest mountain (w_1), or Pollux is on the highest mountain (w_2). His belief set is $Bel_{(w_1,p)} = \{(w_1, p), (w_2, p)\}$. This belief set satisfies Propositionality, and is consistent with Stalnaker's treatment of other examples.[24]

As can be seen in this case, both you and your counterpart god suffer from some ordinary uncertainty – you are both ignorant about which of w_1 and w_2 is the actual world. However, in addition to this, only you lack the *de se* information regarding your own identity, because you do not know whether you are Castor or Pollux. Both your belief sets are plausible and compatible with Stalnaker's account of *de se* uncertainty. However, they happen to be mutually incompatible. This is because, under this analysis, $Bel_{(w_1,c)}$ (your belief set) and $Bel_{(w_1,p)}$ (your counterpart's belief set) have a common element, namely (w_2, p). By Proposition 1, however, if $Bel_{(w_1,c)} \cap Bel_{(w_1,p)}$ is not empty, the two belief sets should be identical – which is not the case here, as they do not coincide on the other elements.

The problem I have just raised appears to originate from the fact that the proper names 'Castor' and 'Pollux' are introduced differently by you and by the other

24 See e.g. the discussion of the parking lot case in Stalnaker (2016). Moss (2012) also makes a similar move in her treatment of *de se* uncertainty.

god. On this version of the story, you have introduced a name via a definite description (e.g. by saying 'the god on the coldest mountain in named Pollux'), while the other god has introduced his own name demonstratively (e.g. by saying 'let myself be named Pollux', where 'myself' contextually picks out the god who utters that sentence). Since the names are introduced differently, we should not expect them to refer necessarily to the same individual. In other words, to correctly pick out the referent of a proper name without ambiguities, we must have an account of how it was introduced. This, however, is a problem for Stalnaker's account. The framework I have presented does not explicitly mention linguistic notions (such as proper names or definite descriptions), but is built around possible worlds and centres. The resulting extensional framework is quite flexible, as in principle it is able to accommodate different ways of individuating objects. This means that both ways of assigning proper names that I have discussed in this section (demonstratively, or through a definite description) can be accommodated within Stalnaker's framework. The problem is that, in cases similar to the example of the two gods, different ways of assigning names can give rise to ambiguities. When that happens it is not clear what proposition is the content of a given utterance. The name 'Pollux', for instance, as we have seen is ambiguous: its referent could either be fixed by a definite description ('the god on the coldest mountain') or demonstratively. The extension of the *de se* belief 'I am Pollux' is a different set of centred worlds depending on how the name Pollux is introduced – it contains the centred world (w_2, p) for one god (the one who fixes the name via a definite description), and the centred worlds (w_1, p) and (w_2, p) for the other god (who demonstratively names himself Pollux). This, as we have seen, leads to a violation of Proposition 1. So, the very flexibility of Stalnaker's framework – the fact that it can formally accommodate different ways of fixing the referent of proper names – leads to inconsistency.

A possible response to the problem I have just raised may be to admit only one way of fixing the reference of proper names – for instance, only allowing proper names to be introduced demonstratively. There are several reasons why, in my opinion, this would not be a satisfactory solution. In particular, as I have noted above, the fact that Stalnaker's framework is formulated in extensional terms makes it in principle very flexible. However, allowing only one way of introducing proper names would preemptively limit the applicability of the framework. Moreover, alternative accounts of the content of *de se* beliefs do not give rise to similar inconsistencies. The Lewisian account, in particular, does not require us to define an epistemic accessibility relation over the set of centred worlds, and it allows overlapping belief sets. This makes it more suitable to accommodate examples such as that of the two gods that I have discussed in this section.

4.5.4 *Updating*

The upshot of the discussion of maximal uncertainty in the previous section brings up another issue for Stalnaker's framework, concerning the way in which it can be used to model how agents update their beliefs over time. As recalled in section 4.2, compatibility with standard accounts of Bayesian updating is commonly regarded as a positive feature of the Weak Acceptance view.[25] However, if what I argued in the previous section is correct, this raises some problems for the applicability of conditionalisation to belief sets.

In order to see where the problem lies, it will be useful to recall briefly how belief updating is standardly formulated in a possible worlds framework, and then examine how this extends to the centred worlds framework put forward

25 Moss (2012). I will have more to say on this topic in Chapter 6.

by Stalnaker. So let's begin by defining the basic terms in a standard possible worlds framework to represent belief. Let W be a set of possible worlds, and $X \subseteq \mathscr{P}(W)$ be a set of propositions over which an agent a distributes her beliefs. Within this framework, propositions are simply viewed as sets of possible worlds. A proposition $x \in X$ is true if and only if the actual world $w_@$ is one of its elements.

Now let B_0 be the set of all the ordinary possible worlds that are elements of some proposition that a believes at stage $t = 0$. When a learns some new evidence, in the form of some proposition $x \in X$ that she now believes to be true, she eliminates all the possibilities that are not consistent with x from her prior belief set, thus making her new belief set at stage $t = 1$ equal to $B_1 \subset B_0$. A similar process takes place when probabilities are involved, which is called conditionalisation. The only difference is that in addition to eliminating the possibilities that are ruled out by the newly acquired evidence, probabilities are redistributed among the surviving possibilities in a way that preserves the ratios among them. But setting probabilities aside for the moment (we will come back to them in Chapter 5), the key feature of the standard account of conditionalisation is that this is an updating strategy which works in the following way: given an initial belief set B_0 and a new piece of evidence, in the form of a learnt proposition $x \in X$, it returns a new belief set $B_1 \subseteq B_0$, that is the subset of the original belief set which includes all the possible worlds that are compatible with x. In other words, on the standard picture, updating on new evidence means refining the set of worlds considered possible, by excluding those that are inconsistent with the evidence.

This feature of the standard account cannot be reproduced entirely under the Weak Acceptance view of *de se* beliefs, at least without some further qualifications. This is because *de se* beliefs typically change over time in a way that is

not compatible with conditionalisation.[26] For example, while I presently hold the *de se* belief that 'I am now in London', I did not hold this belief one month ago, as I was traveling to Paris. The change in my *de se* beliefs between the time when I was in Paris, and the present time when I am in London, is incompatible with conditionalisation because the beliefs I hold now are not a subset of the beliefs I held a month ago. Rather, some of my beliefs simply have been replaced: namely, the belief 'I am in Paris' was replaced by the belief 'I am in London'.

The proponents of Weak Acceptance, however, argue that there is a way to recover the standard notion of conditionalisation, to make it applicable to *de se* beliefs. The idea is that at any given stage, the contents of an agent's beliefs (including all of his or her *de se* beliefs) are always *equivalent* to standard propositions. For example, my current *de se* belief that 'I am in London' is equivalent, according to this view, to some standard proposition such as 'Silvia is in London on the 10th of May'. At the time I was in Paris, I may not have been sure that the proposition 'Silvia is in London on the 10th of May' is, in fact true. But I did regard it as a possibility, so my belief set at the time included possibilities consistent with 'Silvia is in London on the 10th of May' being true. Come the 10th of May, I acquire the evidence to conclude that 'Silvia is in London on the 10th of May' is indeed true, eliminating all other possibilities from my belief set. When viewed in terms of the equivalent standard propositions that are believed by the agent at each time, the change in *de se* beliefs is therefore analogous to updating based on conditionalisation.

On Stalnaker's account, the idea I just described is cashed out more precisely by specifying a procedure to associate *de se* beliefs with standard propositions, or sets of possible worlds. As we have seen, within Stalnaker's framework an agent a's belief set, encoding both the objective and *de se* beliefs held by the

26 I will have more to say on this topic in chapter 6.

agent at any given point, is represented by a set of centred worlds $Bel_{(w_@,c)}$, where $w_@$ is the actual world and c corresponds to a's perspective on $w_@$. The definition of Propositionality (2) entails that each element of $Bel_{(w_@,c)}$ is uniquely associated with a possible world w. This, in turn, ensures that any subset s of $Bel_{(w_@,c)}$ can be mapped to a standard proposition x, that is the set of all the possible worlds w that are associated to some element in s. Call any subsets s of the agent's belief set $Bel_{(w_@,c)}$ a *de se* belief, and let S be the set of all possible *de se* beliefs. Propositionality therefore entails that every *de se* belief has as a content an ordinary proposition, or that there is a function $f : S \rightarrow \mathscr{P}(W)$ that given any *de se* belief $s \in S$, outputs an ordinary proposition $x \subseteq W$, which is the content of s.

This gives us a precise way to state how belief sets can change according to conditionalisation on Stalnaker's framework. Let w_1 be the possible world where 'Silvia is in London on the 10th of May' is true, and w_2 be the possible world where the same proposition is false. Suppose now that at t_0, as I am in Paris, my belief set is $Bel_{(w_1,t_0)} = \{(w_1, t_0), (w_2, t_0)\}$. $Bel_{(w_1,t_0)}$ is equivalent to the standard proposition $y = \{w_1, w_2\}$ ('Either Silvia is in London on the 10th of May, or she is not'). As I progress to t_1, my belief set changes to $Bel_{(w_1,t_1)} = \{(w_1, t_1)\}$, which is equivalent to the proposition $x = \{w_1\}$ ('Silvia is in London on the 10th of May'). The change in my beliefs, with respect to the standard propositions that are mapped to my *de se* beliefs at t_0 and t_1, appears compatible with conditionalisation. If we only look at the propositions that are mapped to my beliefs, at t_0 I believe the equivalent of proposition y, and at t_1 I believe the equivalent of propositions x, which is a subset of y. The additional evidence I have learned between t_0 and t_1 allowed me to rule out the possibility that w_2 is the actual world.

Based on the discussion in the previous paragraphs, it therefore seems that Stalnaker's framework can successfully reproduce conditionalisation in the con-

text of *de se* beliefs. In the remainder of this section, however, I will challenge this solution. In a nutshell, the problem that I wish to raise is that Stalnaker's framework does not explain how agents come to have *de se* beliefs.

Consider again my earlier example, about my current *de se* belief that 'I am in London' (s). Stalnaker's account makes the content of this belief equivalent to a standard proposition, such as 'Silvia is in London on the 10th of May' (x). Stalnaker's argument for saying that I come to believe s by applying conditionalisation rests on the idea that the proposition I come to believe is x, and that –for *me*– x at t_1 implies s.

But how, exactly, does x imply s? We have seen that by the definition of Propositionality, s entails x, because x is the content of s. The converse, however, does not hold in general, because the *de se* information expressed by s is not reducible to ordinary propositional content. First of all, Non-deducibility (1) explicitly excludes that x alone can entail s. And secondly, this can be easily checked by noting how other *de se* beliefs also entail x. For instance, let $s' = \{(w_1, t_0)\}$ also entails x, even though $s \neq s'$. So if the only information that is available amounts to x, this is compatible with (at least) two *de se* possibilities, s and s'.

This leads us to the following dilemma. On the one hand, on Stalnaker's account, the content of any *de se* belief is equivalent to a standard proposition, given the agent's current belief state. Propositionality ensures that, given any belief state $Bel_{(w,c)}$, every *de se* possibility within $Bel_{(w,c)}$ is associated with a unique possible world w. This makes it possible to view the change in beliefs as new information is learned through the process of conditionalising on standard propositions. On the other hand, however, Propositionality (that is the very same condition that enables the recovery of standard conditionalisation for *de se* beliefs) undermines the possibility of *learning de se* beliefs. Moreover,

it also undermines the very reasons that are often cited in favour of updating beliefs by conditionalisation, as I will explain in a moment.

First, let us consider how the framework undermines the very possibility of learning *de se* information. Consider again my earlier example concerning my *de se* belief that 'I am in London', or $s = \{(w_1, t_1)\}$. As we have seen, the content of this *de se* belief corresponds, given my belief set at t_1, to the proposition 'Silvia is in London on the 10th of May', or $x = \{w_1\}$. At time t_0, the same proposition x corresponded to a different *de se* belief, namely $s' = \{(w_1, t_0)\}$, about which I was uncertain, given my belief state at t_0. What changed between t_0 and t_1? I certainly haven't learned that s' is *true*, because by t_1, s' is simply no longer part of my belief set. What I seem to have acquired is a new *de se* belief, s.

This highlights a troubling feature of the framework proposed by Stalnaker: it makes *de se* information *unlearnable*. And this is not just a feature of the example I chose, but a systematic feature of the framework. To see why, consider that in order for an agent to acquire any new beliefs via conditionalisation, it is necessary that they receive some new piece of evidence. Call this new piece of evidence E. The simple fact of receiving E alters the agent's circumstances, and is reflected in a change in the agent's belief set. By Proposition 1, any two belief sets Bel and Bel' must be either identical or disjoint. What this means is that when I am uncertain about some *de se* information, Stalnaker's framework tells me that I cannot learn *that de se* information, because any piece of evidence that I receive would simply change my belief set. All that can happen is that my current belief set is replaced by a different one as I learn an additional piece of evidence about the world, and this leads me to form a different set of *de se* beliefs.

This, however, leads us to a second, more serious worry, which is that Propositionality might actually undermine the standard case for updating beliefs via conditionalisation. Stalnaker's framework has nothing to say about how an agent's belief state Bel changes over time. The way in which the epistemic accessibility relation R is defined induces belief sets corresponding to any of the centred worlds that might correspond to an agent's circumstances. The framework, however, does not say where R comes from.

A standard defence of conditionalisation is that it is a *conservative* way to update one's beliefs upon learning some new information. What happens in Stalnaker's framework, however, is that belief sets are not merely updated, but are completely replaced as the agent's circumstances evolve. The way in which this replacement takes place, moreover, does not appear to be guided by any set of principles, as the definition of R is, as we have seen, exogenous to the framework. This undermines the standard reasons given for conditionalisation, because it weakens the connection between an agent's beliefs at different times.

4.6 A TENSION BETWEEN NON-DEDUCIBILITY AND PROPOSITIONALITY

The objections I raised in the previous section are primarily directed at the details of the framework that Stalnaker proposes to model *de se* beliefs. This raises the question of whether these issues can be fixed by modifying the details of the framework, or if they are instead symptomatic of deeper issues with the Weak Acceptance position on *de se* beliefs. In this section, I argue that the latter is the case. My argument aims to show that the two core assumptions of Weak Acceptance generate a tension that is difficult to resolve without making the position empirically meaningless. The argument that I present in this

section uses only the informal version of the two core assumptions of Weak Acceptance, discussed in Section 4.3:

Assumption 1 (Non-deducibility)**.** *De se* information cannot be deduced from a purely objective representation of the world. In particular, *de se* beliefs cannot be deduced from a set of objective propositions (in the absence of any further premises).

Assumption 2 (Propositionality)**.** Uncertainty about *de se* information is always also uncertainty about the world.

A natural reading of 2 is as a conditional statement: *if* there is *de se* uncertainty at all, *then* there is also objective uncertainty. In other words, Propositionality entails that *de se* uncertainty implies objective uncertainty.

The converse of Propositionality, on the other hand, entails that whenever there is no objective uncertainty about the world, there cannot be any *de se* uncertainty, either. It is easy to check that the converse of Propositionality must hold if Propositionality does. If it didn't, there would be some cases where *de se* uncertainty coexists with perfect information about the world, but this is ruled out by 2.

The reason why proponents of Weak Acceptance introduce Propositionality is that it ensures that *de se* uncertainty can always be reduced to objective uncertainty. Whenever there is uncertainty about some *de se* information, by Propositionality, the uncertain piece of *de se* information is always equivalent to some piece of objective information. But this creates – at least *prima facie* – a tension with Non-deducibility. By Assumption 1, *de se* information cannot be deduced from objective information. So, on the one hand, Propositionality implies that if an agent has all the objective information, they must also have

all the *de se* information. But, on the other hand, Non-deducibility implies that it is left open, when the objective information is known, which *de se* beliefs the agent should have.

While this clearly indicates that there is a tension between the two core assumptions of Weak Acceptance, this does not yet give rise to a contradiction between Propositionality and Non-deducibility. To see why, consider Stalnaker's own example of Boris and Albert. In that example, Stalnaker argues that both agents are certain about the objective facts and, because they share the same context, they both believe all the same propositions. Moreover, neither of them experiences *de se* uncertainty – thus satisfying the condition of Propositionality. However, the *de se* beliefs held by Boris are different from those held by Albert and this, according to Stalnaker, shows that – in accordance with Non-deducibility – their *de se* beliefs are not fixed by objective facts about the world.

However, even if (as Stalnaker's example indicates) Propositionality and Non-deducibility do not outright contradict each other, there are only two possible ways to specify how *de se* and objective information can be related on the Weak Acceptance view, and the first way leads to contradiction, while the second way leads to an unsatisfactory explanation of *de se* beliefs. In the rest of this section, I will now outline both horns of the dilemma.

Consider Boris's *de se* belief that *he himself* is in the kitchen. Could Boris have been uncertain about this *de se* piece of information? Intuitively, the answer to this question seems to be yes. If Boris was uncertain about whether he was in the kitchen or in the living room, for example, then by Propositionality there would also be some objective information which Boris did not know for certain – for instance, he might not have had the objective information that the individual named Boris is in the kitchen. Call this missing objective information

x, and call Boris's *de se* belief that he is in the kitchen s. The question now is: does x imply s for Boris?

Intuitively, again, the answer to whether x implies s *for Boris* appears to be yes. If Boris didn't know s, but was informed that x, he would presumably also come to believe s. However, by Non-deducibility, we already know that x alone does not entail s (and indeed, if the same information x was given to Albert, we would not intuitively expect Albert to come to believe s). So, x entails s only when taken together with some other prior belief that Boris already has (for example, we may think that x together with the prior belief that 'I am Boris' entails s). Call this prior belief (that is necessary to entail s) ξ.

4.6.1 *The first item of de se belief*

Is ξ a belief about the world, or is it a *de se* belief? If ξ is a belief about the world, then if x and ξ together entail s, it means that s can in fact be deduced from a set of purely objective information (since both x and ξ are objective). This would violate Non-deducibility. Therefore, ξ cannot be an objective belief about the world. But if ξ is not objective, then – presumably, unless some further category of beliefs should be presupposed – it must be a *de se* belief. This reasoning highlights the fact that, on the Weak Acceptance view, the equivalence between objective and *de se* information is necessarily mediated by the presence of some background *de se* belief ξ, the presence of which ensures that other *de se* beliefs are always entailed by objective beliefs, and *vice versa*. The background belief ξ itself, however, could not be entailed by *any* objective beliefs. If ξ was entailed by some objective belief y, this would again violate Non-deducibility. And if ξ could be deduced from y together with *some other*

de se belief, χ, then we could ask the same question about χ: is it or is it not entailed by other facts? and so on indefinitely.

The problem, if ξ is not equivalent to any objective fact y, is that Propositionality entails that ξ must be certain. In other words, there can be no uncertainty about ξ, otherwise Propositionality would require that there must be some objective belief y that would entail ξ, because – by Stalnaker's own lights – *de se* uncertainty always entails *de dicto* uncertainty. If ξ must be certain, however, it is unclear that any *de se* belief with informational content (such as 'I am Boris', or any piece of background *de se* information in the examples that I have used in this chapter) could play the role of ξ. As soon as we specify an informational content for ξ, it seems clear that an agent could plausibly not know this informational content.[27] For Stalnaker's account to work, ξ needs to be both indubitable and have a propositional content, so that it can be used in inferences. However, there are instances where an agent could be uncertain of *every* piece of *de se* information that could play this role. For example, if ξ is equal to 'I am Boris' (in the case of our earlier example involving Boris's *de se belief* 'I am in the kitchen'), it is plausible that Boris could, for some reason, doubt that he himself is the individual who is named that way (for instance, he could have forgotten his own name, or be unaware of the name others use to refer to himself).

In order to find a suitable candidate for ξ, Stalnaker (2008, 2016) proposes that we use a token reflexive belief, such as 'I am thinking *this thought*' (ξ^*), where 'this thought' is taken to refer to the token thought that the agent in question has at the moment he or she reflects on his or her beliefs. According to Stalnaker, anyone entertaining ξ^* cannot doubt its truth, and must therefore be certain of ξ^*. Moreover, he argues that ξ^* has some informational content, namely it conveys the information that '*This thought* occurs'. So, ξ^* can be

27 This, I take it, is the same conclusion also reached by Perry (1979).

used to make inferences; for example from ξ^* and '*This thought* is thought by Boris in the kitchen', one can infer that 'I am Boris and I am in the kitchen'.

The appeal to token reflexive beliefs to give a foundation to his account of *de se* beliefs raises some metaphysical issues for Stalnaker. In particular, the appeal to this kind of beliefs commits him to *haecceitism*, which is the view that the world could differ non-qualitatively without differing qualitatively. For example, two possible worlds w, w' could be identical in all qualitative respects, but still be numerically distinct. While Stalnaker would accept this consequence,[28] *haecceitism* is far from being an uncontroversial view. But even leaving these issues aside – as a proper discussion of the metaphysical commitments of Stalnaker's view would go beyond the scope of this chapter – token-reflexive beliefs such as ξ^* still do not provide a solid foundation for Weak Acceptance, due to another set of problems. In a nutshell, the issue is that such beliefs are *indubitable* only in so far as their content is context-*dependent*, but in order to be used in inferences they need to have content that is context-*independent*.

To illustrate, let ξ^* be 'I am thinking this thought'. Intuitively, as Stalnaker points out, ξ^* is certain, as any agent entertaining ξ^* could not doubt that it is true, of themselves at that time, that they are thinking the thought they are thinking. The specific content of ξ^* seems to depend on the context in which it is entertained, as 'this thought' refers to different token-thoughts depending on the specific circumstances – for example, 'this thought' will pick out a different token depending on the identity of the agent thinking the thought, but also the time and place at which the thought is entertained. However, if the content of ξ^* depends on the context, then it cannot be generally used in inferences in the way that is necessary for the Weak Acceptance account to work. To see why, consider again the example of the two gods. Suppose that Castor is uncertain about whether he is the god on the highest mountain, or the god on the cold-

28 Stalnaker defends a version of *haecceitism* in Stalnaker (2008).

est mountain. Now, following Stalnaker's strategy, we can be sure that Castor has the *de se* belief 'I am thinking this thought' (ξ^*). Is it possible for him to now learn that 'this thought' is located on the highest mountain? If he were to learn any new information, the thought he would be then thinking would be a different thought from the one he was thinking at the time he originally entertained the belief ξ^*. Therefore, if the content of ξ^* is partially determined by the context, it does not offer a base for inferences, because while the belief ξ^* is indubitable *at the time it is entertained*, its content isn't fixed across the inference.

To contrast this issue, Stalnaker can require that the content of ξ^* is fixed and does not change with the context. What this means is that once the thought is entertained, 'this thought' picks out a specific token thought – and the reference is fixed for the future, so that 'this thought' always refers to the same token, and thus the content of ξ^* is mapped to an ordinary proposition, which is context-independent. However, this second strategy also runs into problems. While the belief 'I am thinking this thought' is indeed indubitable for any thinker entertaining it, once we map it to a proposition that is not formulated in the present tense – as Stalnaker's solution would require[29] – it loses this aura of indubitability. In other words, as the context changes, the original content of ξ^* is no longer indubitable – Castor could mis-remember what he had thought, or be forgetful, or mistake his own past thoughts for those of the other god. Therefore, even if the content of ξ^* is context-independent – as is required in order to be used in inferences – it cannot provide the link between propositional and *de se* belief because it is not always indubitable.

29 See also Moss (2012).

4.6.2 *Discussion*

In this section, I have raised a general problem for Weak Acceptance. The problem, I argued, stems from the fact that the two core assumptions of Weak Acceptance, namely Non-deducibility of *de se* beliefs, and Propositionality, stand at odds with each other. The only way in which they can be made consistent, is by positing that agents always have some background belief, ξ, which makes it possible to establish a link between the agent's other *de se* beliefs and objective beliefs about the world. The difficulty is that ξ cannot be an objective belief (on pain of contradicting Non-deducibility), but if it is a *de se* belief, then it must be certain. However, this also seems to imply that ξ cannot have informational content, which begs the question of how it can be used to underpin the equivalence of *de se* and objective beliefs. So, if Weak Acceptance is to be rescued, it would be necessary to find a candidate for ξ that is both certain and has informational content. Stalnaker's strategy of using token-reflexive beliefs, such as 'I am thinking this thought' – did not ultimately work.

If Stalnaker's proposal does not succeed in solving the problem I raised in the previous section, then we should consider that Weak Acceptance does not give an adequate account of *de se* beliefs. What are the alternatives? As recalled in Section 4.1, there are two other alternatives to Weak Acceptance. As I already argued against Denial, Strong Acceptance survives as a plausible remaining option. Strong Acceptance endorses Non-deducibility, but it does not endorse Propositionality. This means that there is no requirement, under Strong Acceptance, that *de se* information must always be mapped to some objective information. In particular, there is no requirement that if an agent is certain about his or her own objective beliefs, he or she must not have any *de se* uncertainty. Thus, Strong Acceptance sidesteps the tension I have described between the

two core assumptions of Weak Acceptance, by endorsing the first, bur rejecting the second.

4.7 CONCLUSION

In this chapter, I have considered the Weak Acceptance view of *de se* beliefs. First, in Section 4.1 I presented three possible alternative views about *de se* beliefs, which I named the Denial, Weak Acceptance and Strong Acceptance views. Weak Acceptance has been the main focus for this chapter (Section 4.2). In Section 4.3, I have identified the two core assumptions that underlie Weak Acceptance, and explained why the combination of Non-deducibility and Propositionality makes for a *prima facie* attractive solution to the problem of *de se* beliefs. Section 4.4 was then devoted to a detailed discussion of the framework proposed by Stalnaker to model *de se* beliefs, and raised some objections to this framework. Next, in Section 4.6 I raised a more general objection to Weak Acceptance, that is only presupposed on the tension between its two core assumptions. I showed that there is no way for Weak Acceptance to resolve the tension without either generating an inconsistency between Non-deducibility and Propositionality, or running into the problem that no item of belief can function as the required link between *de se* and ordinary propositional beliefs in all contexts.

Based on the issues discussed in this chapter, I believe that despite the initial attractions Weak Acceptance is not a tenable position and does not offer an adequate solution to the problem of *de se* beliefs. For this reason, Strong Acceptance emerges as the best way to account for *de se* beliefs.

5

CENTRED PROBABILITY

In previous chapters, I have introduced centred worlds as a framework to represent *de se* or self-locating beliefs. This chapter reviews how probabilities can be defined on centred worlds to model uncertain *de se* beliefs, and then considers which interpretations of probability are available when we consider event spaces defined over centred worlds. Intuitively, one might expect that the subjective interpretation of probability would provide the most natural interpretation for probabilities defined over sets of centred worlds. However, as we will see, all the main interpretations of probability currently available in the literature admit an extension to centred probabilities. Moreover, somewhat surprisingly, centred probabilities pose a serious issue to the diachronic aspect of the subjective interpretation.

As we have seen in previous chapters, centred worlds may be used to capture self-locating uncertainty in different contexts, as the following two examples illustrate:

Example 1. Disturbing bell The town where Ann lives has a bell that sounds twice every morning at 3am and at 4am. Every time it sounds, Ann hears it and just goes back to sleep, forgetting all about it.

When Ann wakes up, knowing that the bell sounds twice and that she always forgets about hearing it after going back to sleep, what should her beliefs be regarding what time it is?

Example 2. Map Tom is lost in a new town and can't find his exact position on the map. His surroundings appear to be compatible with two possible locations. What should Tom's beliefs be with respect to his current location?

Both examples present cases where an agent knows the relevant objective properties of the world (represented by the awareness of the bell setup in 1 and by the map in 2), but there are different possible locations (in time or in space) at which they might take themselves to be located. In previous chapters, I have defended the view that self-locating uncertainty is not reducible to non-self-locating uncertainty, and argued that centred worlds are the right framework to model it. Now, I will proceed to define probabilities over centred worlds.

5.1 FORMAL BACKGROUND

This section briefly introduces the formal background to which I will refer in the following sections. I will first outline the axioms of probability given by Kolmogorov (1933), which constitute the standard mathematical treatment of probability. I will then introduce the notion of centred and uncentred events.

Given a non-empty set Ω, which can be taken to represent all relevant possibilities, an algebra S on Ω is a set of subsets of Ω, which contains Ω and is closed under complementation and union. For example, if $\Omega = \{x, y, z\}$, then $S = \{\{x\}, \{y, z\}, \{x, y, z\}, \emptyset\}$ is an algebra on Ω.

Kolmogorov defines probability as a function p from S to the real numbers, which obeys three axioms:

Axiom 1 (Non-negativity)**.** For all $A \in S$, $p(A) \geq 0$.

Axiom 2 (Normalisation)**.** $p(\Omega) = 1$.

Axiom 3 (Additivity)**.** For all $A, B \in S$ such that $A \cap B = \emptyset$, $p(A \cup B) = p(A) + p(B)$.

In addition to Non-negativity, Normalisation and Additivity, Kolmogorov gives the following definition of the conditional probability of an event B, given another event A:

Definition 5 (Conditional probability)**.** For all B, $A \in S$, $p(B|A) = \frac{p(B \cap A)}{p(A)}$ (assuming $p(A) > 0$).

This definition of conditional probability, also known as the *Ratio formula* for conditional probability, is considered standard and will be assumed in this chapter.[30] Importantly, the ratio formula for conditional probability is not applicable when $p(A) = 0$, in which case the conditional probability of any event given A is left undefined.

In applications of probability theory, the set of possibilities can be identified with different objects depending on the context. In the case of 'ordinary' possible worlds, this coincides with the set W of all possible worlds. An event $E \subseteq W$ is defined as a set of possible worlds. Since the set of all events S forms an algebra on W in the sense explained above, a probability function obeying the three probability axioms can be defined on S.

30 Some authors have been critical of this definition of conditional probability, arguing instead that conditional probability should be taken as a primitive binary function. (See Popper, 1959a; Hájek, 2003).

Centred worlds represent an extension of ordinary possible worlds in the following sense. In addition to encoding information about the world, a centred world picks out a specific location within that world. Given the set of possible worlds W and a set of all possible centres C, a centred world (w, c) is defined as an ordered pair of a possible world $w \in W$ and a centre $c \in C$. Based on this definition, the set of all possible centred worlds, which will be denoted Ω, is a subset of the Cartesian product $W \times C$.[31]

A centred event E^* is defined as a set of centred worlds and S^* is the set of all centred events. Since S^* forms an algebra on Ω, a probability function can be defined on S^*, analogously to the previous case.

In what follows, we will be interested in the relationship between probabilities defined on centred and uncentred (or 'ordinary') possible worlds. Centred events represent more fine-grained possibilities than events defined on ordinary possible worlds, which are not able to capture information about specific locations within a world or set of worlds. The following condition expresses an important relationship between centred and ordinary events (that is, events defined on the set W of ordinary possible worlds):

Definition 6 (Correspondence)**.** For any ordinary event $E \subseteq W$, we can define the corresponding centred event $E^* = E \times C$, which is equal to the set of centred worlds $(w, c) \in \Omega$ such that $w \in E$ and $c \in C$.

Correspondence ensures that for every ordinary event $E \in W$, there exists a centred event $E^* \in \Omega$ that has the property that it contains all the centred worlds which coincide on w, for all $w \in E$. When this is the case, we'll say that E^* *corresponds* to E. The converse does not hold, as there may be several centred events that do not correspond to any uncentred event.

31 Note that since not all centres may be present within each possible world, this definition does not require that all possible combinations of a world and a centre belong to Ω.

Building on definition 6, we can now also define the class of uncentred events defined on the finer-grained set Ω:

Definition 7 (Uncentred event)**.** $E^* \in \Omega$ is *uncentred* if and only if there is $E \in W$ such that E^* corresponds to E.

In other words, according to definition 7, an uncentred event is just a centred event (that is, an event defined on Ω) that corresponds (in the sense set out by definition 2) to an ordinary event. Another type of centred events that will be of interest to us is what I will call an *indexical* event:

Definition 8 (Indexical event)**.** $E^* \in \Omega$ is *indexical* if for all $w, w' \in W$ and for all $c \in C$ such that $(w, c), (w', c) \in \Omega$, $(w, c) \in E^*$ if and only if $(w', c) \in E^*$.

Whereas an uncentred event represents a condition of maximal uncertainty with respect to c, an indexical event corresponds to a condition of maximal uncertainty with respect to w.

5.2 INTERPRETATIONS OF PROBABILITY

Imagine a coin that is just about to be tossed. Intuitively, we may think that the probability that the coin will come up Heads is 1/2. But what exactly does this statement mean? In other words, what *is* a probability and *how* can it be assigned to specific events? Different interpretations of probability offer different answers to these basic questions. This section gives a survey of the main interpretations that have been put forward in the literature, which fall into three main types: logical, objective, and subjective. But before introducing them, it will be useful to state some criteria that can be used to evaluate their adequacy.

5.2.1 *Criteria of adequacy*

Following Hájek (2012), we can identify three broad criteria of adequacy for an interpretation of probability.

The first criterion is *Admissibility*. According to this criterion any proposed interpretation of probability should be compatible with the mathematical treatment of probability, which was reviewed in the last section. If a proposed interpretation of probability is not compatible with the three Kolmogorov axioms of probability, we will say that it is *inadmissible*.

The second criterion is *Ascertainability*. This requires that an interpretation should admit that we could, at least in principle, find out what the correct probabilities are. Hájek points out that specifying what 'in principle' means could be quite tricky; we will set this issue aside for the moment.

Finally, *Applicability* is the requirement that an interpretation of probability should be useful to explain our intuitive concept of probability in different areas where it seems appropriate to apply it. Firstly, it should make sense of the intuition that some events have intermediate probability values. Secondly, probabilities are used in a wide variety of contexts: among other things, they can be used to describe the likelihood of precipitation in London tomorrow afternoon, the chance of drawing an Ace from a deck of 54 cards, the eventuality that an uranium atom will decay within a certain time, how likely it is that Jones will catch a cold after walking in the rain all afternoon, the statistical incidence of car accidents on weekends, and so on.

All these examples pick out seemingly very different phenomena. An adequate interpretation of probability should explain how probabilities relate to each of these. For Hájek, this idea could be formulated as a cluster of sub-criteria:

applicability to frequencies (it should inform our understanding of the relationship between probabilities and frequencies); applicability to rational belief (it should say something about the relationship between probabilities and rational beliefs); applicability to ampliative inference (it should explain the role played by probabilities in evidential support relationships); and applicability to science (it should explain how probabilities figure in scientific theories).

5.2.2 *Logical probability*

Logical interpretations view probabilities as proportions within a possibility space. Probabilities can be assigned to events in the absence of empirical information, as they basically encode logical or evidential support relationships that exist between events within a given space of possibilities. Each basic possibility w_i is assigned a mass of value $m(w_i)$, such that $m(w_i) > 0$ for at least some $w_i \in W$. The function m will be called a *probability mass function*.

A probability mass function only assigns a weight to the basic possibilities, so to get a fully fledged probability function we need to extend it to all the possible events. To achieve this, note that m induces a function m^* on the set of all events $E \subseteq W$ which, according to the logical interpretation, represents the probability of E. To be more precise, the probability $p(E)$ of an event $E \subseteq W$, which is equal to $m^*(E)$, is given by the proportion between the sum total of m-weighted elements of E and the sum total of m-weighted elements of W.[32] Formally:

32 In this chapter, I am restricting the discussion to finite possibility spaces, i.e. where the set W is finite. However, it should be noted that summation does not work when W is uncountably infinite. In that case, we need to integrate rather than add the basic possibilities, or use measures in general.

Definition 9 (Logical probability)**.**

$$p(E) = m^*(E) = \frac{\sum_{w_i \in E} m(w_i)}{\sum_{w_j \in W} m(w_j)}.$$

So defined, m^* satisfies all three axioms of probability and provides an admissible interpretation of probability.

In the absence of external constraints, there are in principle an infinite number of possible specifications of the weight function m. The two main Logical interpretations of probability differ in the way they specify how to pick a specific m.

According to what is known as the Classical interpretation,[33] if all the elements w_i of W appear *equally possible*, or in other words if we don't have any reason to discriminate between them, they should receive an equal weight.[34] In particular, it seems natural to stipulate that each element w of W has a weight of 1. The value of the weighted sums in Definition 5 then becomes equivalent to the cardinalities of E and W. Definition 6 can therefore be simplified for the Classical interpretation as follows:

Definition 10 (Classical probability)**.**

$$m^*(E) = \frac{|E|}{|W|}.$$

We can illustrate how the Classical interpretation works with our simple coin toss example. Here, there seem to be two possible cases: either the coin toss

33 The name derives from the fact that it was historically the first to be proposed. A formulation of this interpretation of probability can be found in the work of Laplace.

34 In the case of a finite W, this idea is expressed by what is known as the *Principle of Indifference* (PI), which tells us that when we have no external reasons to discriminate between all the elements of W, we should assign equal probabilities to each of them. The same idea is generalised to the case where W is countably infinite by the Principle of Maximum Entropy.

comes up Heads, or it comes up Tails. Since both cases are possible and there is no apparent reason to discriminate between them, the Classical interpretation of probability tells us that they should receive an equal weight. By applying the definition of Classical probability, we then get $m^*(Heads) = \frac{|\{Heads\}|}{|\{Heads, Tails\}|} = \frac{1}{2}$.

The Classical interpretation seems to work well in our simple coin toss example, where it delivers the intuitively right result. However, in some situations it is not possible, or at least intuitively plausible, to assign equal weights to all the basic possibilities. To illustrate this point, we can go back to our coin toss example, but consider now the additional possibility that the coin toss lands on the edge and amend our model by defining W as containing three elements, namely $Heads$, $Tails$ and $Edge$. Since these are clearly distinct possibilities, a quick application of the Classical interpretation would assign an equal weight to each. However, at least if we assume that the coin is tossed under normal conditions, it clearly seems that $Edge$ is extremely unlikely to occur, so we wouldn't be justified in treating it as equal to $Heads$ and $Tails$.

This simple example highlights a problem for the Classical interpretation of probability: whenever an equal assignment of weights to the elements of W does not appear to be justified, the Classical interpretation cannot be reasonably applied. More generally, the probabilities assigned by the Classical interpretation will not be invariant under redescriptions of the underlying basic cases.[35]

This problem of applicability is solved by what I will call Generalised Logical interpretations of probability,[36] which allow the weight function m to assign unequal weights to the elements of W.

35 For simplicity, the discussion of this chapter is limited to the case where W is finite. However, the points made remain valid when the space of possibilities is infinite. When W is not finite, additional problems of underdetermination for the Logical interpretation arise, as demonstrated by the well-known 'Bertrand paradox' cases.

36 This is usually called the Logical interpretation. I add the qualifier 'Generalised' to distinguish it from the class of interpretations under which it falls.

As noted above, there are still an infinite number of candidates for m and each specification of m gives rise to a different Generalised Logical interpretation. Any choice of m ought to be justified on the basis of some principle, which sets out an *a priori* method to determine $m(w_i)$ for all the elements of W. Carnap (1950), the main proponent of this type of interpretation of probability, gives some principles of regularity based on the symmetries that can be traced in the linguistic descriptions of the elements of W. This strategy, however, gives rise to some worries. In particular, the resulting probability assignments will vary depending on the particular choice of language used to formulate the descriptions, limiting the applicability of this type of interpretations. If the principles and the language chosen to formulate the descriptions are not independently justified, the risk is that any choice of a particular m would be arbitrary.

In conclusion, according to Logical interpretations, probabilities can be determined *a priori* from the structure of the set of all possibilities W. The probability of an event $E \subseteq W$ is given by a weighted proportion between the cases included in E and the total cases in W, as specified by Definition 9. How do Logical interpretations fare with respect to the criteria of adequacy outlined in §5.2.1? Both the Classical and the Generalised Logical interpretations are admissible and ascertainable, as long as the set W is given and m is correctly specified. However, they appear to raise some issues with respect to the criterion of applicability. In the case of the Classical interpretation, the problems arise with the fact that it is not always reasonable to assign equal weights to the elements of W, and that probability assignments given by this interpretation are not invariant under redescriptions of the basic cases. In the case of the Generalised Logical interpretations, the main problem seems to be the arbitrariness in the specification of m.

5.2.3 *Objective probability*

According to Objective interpretations of probability, the probability of an event E is just the objective chance that E will happen. Objective chances, in turn, are facts about the world that are fixed independently of anyone believing them and the language used to model a specific situation. This characteristic conceptually sets Objective probabilities apart from the Subjective and Logical interpretations.

Formally, we can represent objective chances as a function ch that assigns to each event $E \subseteq W$ a real number. Objective interpretations make the claim that the probability of an event E will just be equal to its objective chance; formally:

Definition 11 (Objective probability)**.**

$$p(E) = ch(E).$$

The above definition places no constraint on the specification of ch and is therefore not very informative until something more is said about what ch is and how it can be determined. Different types of Objective interpretations differ in the way that they handle this task.

The Frequentist interpretation of probability identifies the objective chance of an event E with the relative frequency with which E occurs within a suitable reference class. To illustrate how this works, we can again go back to our simple coin toss example, and imagine that we are about to perform a toss. What is the chance that it will come up Heads? This chance experiment involves two possible outcomes, so we can model it using the set of possible worlds $W = \{Heads, Tails\}$. According to the Frequency interpretation, $ch(Heads)$

should be determined relative to a reference class. A natural way to identify the appropriate reference class in this case is to take a series of runs of the same type of experiment, that is a sequence of coin tosses.[37]

At this point, there are two ways in which a frequentist could go. A first possibility is to take as the reference class an actual sequence of observed coin tosses (see Venn, 1876). This strategy has the advantage of being admissible and empirically ascertainable, but it generates some problematic results. One problem is that actual frequencies can only be defined for event types (that is, events that could be multiply realised over time: in our example, the type of coin toss experiments), but often times we want to give chances to singular – or *token* – events (in our example, we might be interested in the chance that a particular coin toss will result in heads). This is known as the 'problem of the single case'. Another problem is that even if we are only dealing with type events, the probability of a type event that will only start happening in the future is undefined. For example, we might want to give the chance that coin tosses performed with a coin that will be minted tomorrow will result in heads. We cannot presently give any chance to this event type, because no instances have been recorded yet.

Another possibility for the frequentist is to take the limiting frequency of an event within a hypothetical reference class. On this approach, the objective chance of Heads is identified with the limiting frequency of its occurrence if the coin was tossed infinitely many times. Since the reference class in this case is fixed independently of the observed coin tosses, this approach appears to solve the problem of the single case (see Reichenbach, 1949; von Mises, 1957).

37 Strictly speaking, possible worlds would have to represent complete histories, which are sequences of coin tosses in our simple example. In order to identify possible worlds with outcomes of individual coin tosses, I am making an important simplification here, treating each repetition of the coin-tossing experiment as a complete 'history'. The collection of all these individual histories constitutes the reference class for the probability of the coin toss landing Heads.

However, hypothetical limiting frequencies also raise issues on other levels. Probably the most relevant one is what is known as the 'problem of the initial sequence'. When working with hypothetical frequencies, the reference class is an infinite series of coin tosses. But this seems to place too little constraints on actual observations of repeated coin tosses: assuming that the coin is not completely biased against Heads or Tails (so that both outcomes have a positive chance of occurring, of whatever value), then any observed sequence of outcomes is compatible with any assignment of probability to Heads. Suppose that we tossed the coin one hundred times, and it landed Heads exactly 53 times. This observation would in principle be consistent with the coin having an objective chance $ch(Heads) = .5$ of landing Heads, but it would also be consistent with $ch(Heads) = .6$, or even $ch(Heads) = .1$.

To see this, consider that the hypothetical reference class includes infinite coin tosses. If there is a positive probability that the coin lands Heads on some of the trials, then there will be an infinite number of trials on which it lands Heads within that reference class. When a subset of the infinite reference class is observed, it could in principle come from any section of the infinite hypothetical sequence. Since we should expect any possible combination of Heads and Tails to occur at some point within the infinite hypothetical sequence, the observation would not place any constraint on the limiting relative frequency of Heads versus Tails. In other words, the 'initial sequence' that is observed is compatible with any assignment of probabilities to Heads and Tails; empirical observations do not place any constraints on the hypothetical limiting frequencies. On the hypothetical frequency approach, therefore, the objective chance of an event seems to be independent of what might intuitively be seen as relevant empirical observations.

Hypothetical frequencies pass the test of applicability and admissibility.[38] Due to the problem of the initial sequence, hypothetical frequencies are not ascertainable, because on this approach empirical observations almost never constrain objective chances.

For the reasons I have mentioned, it seems problematic to identify objective chances with frequencies (either finite or limiting). However, this is not to say that frequencies bear no relationship to probabilities and objective chances in particular. Even if we accepted that relative frequencies do not *constitute* objective probabilities, they could still be taken to *indicate* them.

Another explanation of the nature of objective chance is given by what is called the Propensity interpretation of probability (Popper, 1959b). According to this interpretation, ch expresses the intrinsic tendency (or propensity) of a given situation or physical setup to give rise to a certain outcome. In the coin toss example, $ch(Heads)$ equals the tendency of the specific coin, tossed under normal conditions, to land on Heads.

The Propensity interpretation seems to give a meaningful answer to the question of what is objective chance, but the worry is that it is doing so by positing a metaphysically mysterious property. The main problem with the Propensity interpretation is that it is unclear just what propensities are and how they can be determined, which seems to limit the applicability of this interpretation. If propensities are a type of physical property, they should be measurable, but it is unclear what sort of instrument could be used to accomplish this task. This is problematic with respect to the ascertainability criterion.

At this point, the strengths and weaknesses of the Frequency and the Propensity interpretations may appear to be in some sense complementary. The Propen-

38 To be precise, since hypothetical limiting frequencies are defined on infinite reference classes, in order to pass the test for admissibility Axiom 3 (Additivity) needs to be reformulated to comprise *countable* additivity.

sity interpretation gives some content to objective chances, but it doesn't provide a clear indication of how to determine them in practice. Relative frequencies, on the other hand, don't seem to be a good candidates to be identified with objective chances, but they do seem to be a manifestation of them and could therefore be taken to indicate or estimate the underlying objective chances. A third type of Objective interpretation of probability, the so-called Best-system interpretation, to some extent combines these insights.

According to the Best-system interpretation, objective chances are determined by the laws of nature. In other words, $ch(E)$ is the probability assigned to E by the laws of nature, which in turn are the theorems of the best scientific theory available to us. With this in mind, a proponent of the Best-system interpretation may adopt a functionalist view about objective chance. On this view, there are some properties that characterise objective chance; a probability function p therefore counts as objective chance just in case it satisfies those properties.

List and Pivato (2015), building on Shaffer (2007) and Glynn (2010), offer a list of six properties that characterise objective chance and show that these are satisfiable within a framework where probabilities are assigned to ordinary possible worlds.[39] These six properties concern the relationship between objective chance and other key properties, linking it to precise notions of possibility, contingency, belief, time, causation, the laws of nature and the intrinsic properties of a given context.

This functionalist approach makes objective chance both admissible and applicable. There is some controversy about how to determine what scientific theory is the best, which seems to pose an issue with respect to the criterion of ascertainability. But setting this problem aside, the Best-system interpretation, with the functionalist approach, appears successful in solving the prob-

39 In their framework, possible worlds are equivalent to complete histories.

lems that affect the frequency and propensity interpretations, while retaining their main features. In particular, it makes sense of the relationship between relative frequencies and chances, and can be used to explain the link between Objective and Subjective probabilities.

5.2.4 *Subjective probability*

While the Logical and Objective interpretations view probabilities as properties of events, the subjective interpretation is concerned with the degrees of belief that agents assign to propositions corresponding to different events. On a Subjective interpretation, the probability of an event E is the degree of belief or credence $cr(E)$ that an agent i has in the possibility that E occurs. Formally:

Definition 12 (Subjective probability)**.**

$$p(E) = cr_i(E).$$

As in the case of ch in the context of Objective interpretations, Definition 8 is relatively broad and needs to be supplemented with an account of cr that specifies its properties and how it can be determined. In particular, we will need to specify the agent who holds the beliefs and the characteristics of those beliefs. There are different analyses of Subjective probability that vary according to how they specify both points.

One approach is to take the beliefs of an actual agent. On this *unconstrained subjectivist* approach, the probability of an event E, for an actual agent i should be understood as the actual degree of belief that i assigns to E. There are various ways in which one could try to implement this approach: for example, one

might identify an agent's degree of belief on a proposition p with the odds that they would be prepared to take for a bet on p.

A first problem with the unconstrained subjectivist view is that there is no guarantee that cr will be admissible. In fact, most of the time actual agents hold beliefs that could not be mathematically represented by a probability function, because they violate the axioms of probability (see Staffel, 2015). Moreover, even assuming that the credences of actual agents were somehow ascertainable and admissible, this approach does not seem to offer much in terms of applicability: probability assignments will always be descriptive claims that are relative to actual agents, with no criterion to asses their applicability to other domains.

A second approach solves this problem by taking the beliefs of a *rational* agent. According to the $Bayesian$ version of subjective probability, this means two things in particular: that the agent's credences should conform to the probability calculus; and that they should be updated in accordance to a specific rule over time as new information is learned.

The first thesis places a synchronic constraint on cr, and it is known as *probabilism*. Since probabilism does not seem to be a descriptively accurate thesis, the arguments usually given by Bayesians to justify it point to its normative import.[40] Pragmatic arguments establish that if cr does not conform to the probability calculus, the agent is liable to behave irrationally. In particular, Dutch Book arguments show that if an agent has non-probabilistic credences, he will be disposed to accept as fair a set of bets that guarantees him a sure loss (see de Finetti, 1937; Hájek, 2008). Moreover, representation theorems show that if an agent exhibits coherent preferences between actions that have uncertain con-

40 See (Titelbaum, 2016a) for a discussion of justifications of probabilism.

sequences, his credences can be represented by a probability function (Savage (1954); Ramsey (1931); Jeffrey (1983)).

In addition to the pragmatic arguments, probabilism has also been defended on purely epistemic grounds using *accuracy* arguments. The aim of credences seems to be to approximate as closely as possible the truth; if we measure their accuracy using a class of particularly appealing rules, called 'proper' scoring rules, Joyce (1998) has shown that only credences that obey the probability calculus maximise accuracy.

The second Bayesian thesis places a diachronic constraint on how cr changes over time, and the rule that is generally advocated is known as *conditionalisation*. This rule states that the agent's degree of belief in event E at time t_2, after learning a proposition X, should be equal to her conditional degree of belief in E given X at t_1 (provided that at t_1 the agent's degree of belief in E is not 0). Formally:

Definition 13 (Conditionalisation)**.**

$$cr_{i,t_2}(E) = cr_{i,t_1}(E|X).$$

The arguments offered to justify conditionalisation are similar in spirit to the ones for probabilism. 'Dutch Strategy' arguments show that, if an agent does not update his beliefs via conditionalisation, he is vulnerable to accepting a series of individual bets which he will judge as fair in succession, but which will guarantee him a sure loss on the whole (see Lewis, 1999).[41]

'Orthodox' Bayesians place no further constrains on cr apart from probabilism and conditionalisation. This means that agents who share the same evidence

41 An epistemic argument for conditionalisation based on accuracy is also given by Greaves and Wallace (2006).

about the world could in principle have different credence functions. As long as each has probabilistic credences and updates them via conditionalisation, they all represent valid interpretations of probability.

So-called 'objective' Bayesians, however, have pointed out that epistemically rational beliefs should intuitively satisfy some additional desiderata. In particular, they should match objective chances, if they exist. Lewis (1980) argues that for any event $E \in W$, $cr(E)$ should match $ch(E)$, whenever the latter is known and not trumped by what he calls 'inadmissible' evidence (that is evidence that is relevant to E but does not bear on its objective chance). This constraint is captured by the following:

Definition 14 (Principal Principle)**.** $cr(E|ch(E) = x) = x$.

Does the subjective interpretation of probability pass the three criteria of adequacy? The unconstrained subjective version, as we have seen, does not pass either the admissibility or the applicability criterion. The Bayesian versions, on the other hand, are admissible (because they accept probabilism) and at least in principle ascertainable, as cr could be derived from choice behaviour (via representation theorems) or through introspection. They also generally meet the applicability criterion, as evidenced by the fruitful ways in which Bayesianism has been applied to many different fields, from statistics to decision theory.

5.3 INTERPRETATIONS OF CENTRED PROBABILITY

In the previous section, I have given an overview of the interpretations of probability that are generally discussed in the literature. These interpretations were designed to work within a possible worlds framework, that is when Ω is identified with the set W of all possible worlds. However, as explained in section 1,

this framework could be enriched by considering not only possible worlds, but also centres within them. In this case, Ω should be identified with the set of all possible centred worlds, defined as ordered couples of a possible world and a centre within it.

In the cases presented in examples 1 and 2 in the opening of this chapter, an agent has full knowledge of the relevant properties of the world (represented by the awareness of the bell setup in example 1 and the map in example 2), but there are two possible locations (in time or in space) at which the agent might be. Let $W = \{w\}$ be the set of possible worlds and $C = \{3am, 4am\}$ the set of possible centres. In example 1, the set of centred possibilities Ω contains the two centred worlds $(w, 3am)$ and $(w, 4am)$ (where w represents the known facts about the bell and Ann's propensity to wake up and forget). In example 2, let $W = \{w\}$ and $C = \{x, y\}$. In this case, Ω also contains two centred possibilities (w, x) and (w, y) (where w represents all the information recorded in the map). On this representation, the uncertainty faced by Ann and Tom in examples 1 and 2 only concerns their own location (temporal or spatial) within the world, but not what the world is like. In other words, Ann's and Tom's situations cannot be captured by *uncentred* events as set out in definition 3.

Uncertainty of location can sometimes be mixed with uncertainty about the world, as illustrated by the following (which modify examples 1 and 2):

Example 3. Uncertain bell The town where Ann lives has a bell that sounds once or twice every morning, depending on the toss of a fair coin. If the coin lands Heads, the bell only sounds once at 3am. If the coin lands Tails, the bell sounds twice, at 3am and at 4am. Every time the bell sounds, Ann hears it and just goes back to sleep, forgetting all about it.[42]

42 This example is tructurally similar to the Sleeping Beauty case, which I discuss in detail in Chapter 7.

Example 4. Uncertain map Tom is lost in a new town. He has two maps that cover different areas, but he is not sure which one represents the town he's in. His surroundings appear to correspond to a specific location on one map, but are compatible with two distinct locations on the other map.

In examples 3 and 4, Ann and Tom, in addition to being uncertain about their location within a world, are also uncertain about what the actual world is. Ann doesn't know whether it is 3am or 4am and, in addition, she is uncertain whether the bell is set to sound once or twice today. We can represent this by expanding the set of possibilities. Now $\Omega = \{(w_T, 3am), (w_T, 4am), (w_H, 3am)\}$ (where w_T and w_H correspond to the possible worlds where the coin toss is Tails or Heads).

In example 4, Tom doesn't know which map is the correct one and is also uncertain of his own position. Analogously with example 3, we can represent the new set of centred possibilities as $\Omega = \{(w_1, z), (w_2, x), (w_2, y)\}$, where x, y and z are the possible locations and w_1 and w_2 correspond to the two maps.

5.3.1 *Criteria of adequacy*

From a technical point of view, a probability function satisfying the Kolmogorov axioms can be defined on centred worlds in much the same way as illustrated in §5.1. This ensures that the admissibility criterion can still be satisfied by probabilities defined on centred worlds. The other two criteria of ascertainability and applicability also carry over from the previous discussion.

However, I will also introduce a fourth criterion of adequacy for interpretations of probability meant to be applied to centred worlds, which I will call *Compatibility*. The compatibility criterion is meant to provide a link between

the probabilities of centred and uncentred events. Intuitively, what it says is that the probability assigned to an *uncentred* event by a probability function p defined over centred worlds should always match the probability assigned to the 'same' event by the coarser probability function p' defined over standard possible worlds. Using the definition of correspondence (definition 6 in §5.1), we can formally state the compatibility requirement. Let p^* be the probability function on the finer-grained (centred) algebra S^* and p be the probability function on the coarser-grained (ordinary) algebra S. Then:

Definition 15 (Compatibility)**.** For every $E \in S$ and $E^* \in S^*$, $p(E) = p^*(E^*)$, whenever E^* corresponds to E.

Compatibility places an intuitively plausible formal constraint on probabilities defined over centred worlds. If centred worlds represent an extension (finer-grained) of possible worlds, we would expect the probabilities to harmonise. If an interpretation of probability can be applied to centred worlds, then it should deliver the same assignment to the corresponding uncentred events when applied at the two levels.

5.4 LOGICAL INTERPRETATION

Starting from this section, I will now go through each of the interpretations of probability that I have outlined in the previous sections, and consider whether it can be meaningfully extended to probabilities defined over centred events.

The Logical interpretation of probability is designed to assign probabilities to events in the absence of empirical information. As explained in section 5.2.2, the main tenet of the Logical interpretation is the principle of insufficient reason. When two possibilities are indistinguishable from the point of view of

all relevant properties, they should receive an equal probability. The Classical interpretation is a special case.

The Classical interpretation seems applicable to centred worlds. Take example 1, for instance. Here, Ω contains two distinct centred possibilities, $(w, 3am)$ and $(w, 4am)$, which appear equally plausible. Following the Classical interpretation, we could therefore assign an equal probability of $1/2$ to each, which also seems the natural solution. Analogously, in example 2 we could assign an equal probability of $1/2$ to the two centred possibilities (w, x) and (w, y) that represent the alternative locations at which Tom might be. In both of these examples, an application of the Classical interpretation seems intuitively plausible.

As discussed in section 5.2.2 regarding the Logical interpretation for possible worlds, the Logical interpretation passes the test for the admissibility and ascertainability criteria, but has some problems with the applicability criterion, which will carry over when it is applied to centred worlds. With respect to the new compatibility criterion, however, it does not fare very well, as we will see in a moment.

Consider the representation of example 4 and the probabilities that the Classical interpretation might assign to the three centred possibilities (w_1, x), (w_1, y), (w_2, z). At a first blush, all three appear to be equally plausible; the Classical interpretation might therefore assign an equal probability of $1/3$ to each of them. Under this description the two uncentred events $E_1 = \{(w_1, x), (w_1, y)\}$ (corresponding to the first map being the right one) and $E_2 = \{(w_2, z)\}$ (corresponding to the second map being right) receive different probabilities, with $p(E_1) = 2/3$ and $p(E_2) = 1/3$. In other words, applying the Classical interpretation to the centred worlds representation of example 4 entails that map 1 is twice as likely to be the correct map as map 2.

According to the compatibility criterion, the probability p of an uncentred event within a centred world representation should match the probability p^* of the same event within the coarser possible-worlds representation. As the number of possible worlds in example 4 is two, corresponding to w_1 and w_2, this means that $p(E_1)$ should be equal to $p^*(\{w_1\})$ and $p(E_2)$ should be equal to $p^*(\{w_2\})$. However, applying the Classical interpretation to $\Omega^* = \{w_1, w_2\}$ yields $p^*(\{w_1\}) = p^*(\{w_2\}) = 1/2$. The application of the Classical interpretation to the centred-world level of description is therefore not compatible with its application at the possible-worlds level of description.

Two replies could be offered to meet this problem. Firstly, one may try to argue that the problem only arises for the Classical interpretation, but can successfully be dealt with by different specifications of the Logical interpretation. Secondly, one may also argue against the compatibility criterion, on the grounds that there is no good reason to expect that probabilities defined on different levels of description would match in any relevant way. I'll consider each of these two strategies in turn.

5.4.1 *The generalised Logical interpretation and compatibility*

To pass the compatibility criterion, a Logical interpretation should ensure that the probabilities assigned to uncentred events match the probabilities assigned to the corresponding possible worlds in the coarser representation. To get this result, a generalised Logical interpretation could simply postulate that the probability assigned to each centred possibility (w_i, c_j) should be inversely proportional to the total number of centred possibilities that share the same w_i. In the case of example 4, this strategy would yield the following assignment: $p^*((w_1, x)) = p^*((w_2, y)) = 1/4$, $p^*((w_2, z)) = 1/2$. Given this assignment,

the probabilities of the two uncentred events E_1^* and E_2^* come out equal, and they match the probabilities that are assigned to the corresponding possible worlds.

Unfortunately, this strategy runs a risk of circularity: the only reason to impose this requirement seems to be that it gets the desired result. Moreover, one could note that at the finer level of description, one could also place a corresponding requirement that the probability assigned to a centred world be inversely proportional to the total number of centred worlds that coincide on c_j. In other words, we could require that probabilities be equally distributed across the partition induced by the indexical events. In the context of example 4, these are $E_3^* = \{(w_1, x)\}$, $E_4^* = \{(w_2, y)\}$ and $E_4^* = \{(w_2, z)\}$. From the perspective of the finer-grained set of centred possibilities Ω^*, this second criterion would seem to be at least as justified as the one introduced before.

However, the second criterion does not satisfy the compatibility requirement. To see this, consider that following the second criterion in the case of example 4 would yield a probability assignment that is equivalent with the Classical interpretation, that is $p^*((w_1, x)) = p^*((w_2, y)) = p^*((w_2, z)) = 1/3$. But we have seen that this probability assignment violates the compatibility criterion. Moreover, the two criteria are in general both mutually incompatible and incompatible with the Classical interpretation, as can be illustrated using Example 3. In that example, Ann faces three centred possibilities, $(w_T, 3am)$, $(w_T, 4am)$ and $(w_H, 3am)$, two of which coincide on the centre (in this case, the time point of 4am). Following the second criterion, a generalised Logical interpretation might assign the following probabilities: $p^*((w_T, 3am)) = p^*((w_H, 3am)) = 1/4$, $p^*((w_T, 4am)) = 1/2$. In contrast, following the first criterion would have yielded $p^*((w_T, 3am)) = p^*((w_T, 4am)) = 1/4$, $p^*((w_H, 3am)) = 1/2$.

5.4.2 *Objections to the compatibility requirement*

What reason is there to expect that the probabilities assigned to events under different levels of description would match? After all, the Logical interpretation of probability is designed to work with a specified language. When the description of the relevant space of possibilities is altered, this makes it impossible to draw meaningful links between the probability assignments that would arise in each case.

This objection to the compatibility criterion deserves to be taken seriously, but there are still strong pragmatical reasons to think that the compatibility criterion should be kept. It would be very impractical for any application of probability if we could not assume that an interpretation of probability would deliver the same results when applied at different levels of description, when the finer level of description does not add any new information. Centred events, defined at the finer grained level of description corresponding to the set of centred possibilities Ω, in general convey more information than the events defined at the coarser grained level of description Ω^*, because they contain information about the location within a possible world that is not expressible at the level of Ω^*. However, the uncentred events at the level of Ω are by construction equivalent to the coarser grained events at the level of Ω^*, so the former should be taken to be informationally equivalent to the latter.

5.5 OBJECTIVE INTERPRETATION

According to Objective interpretations, probabilities are physical facts determined by the laws of nature. On a first blush, it would therefore seem impos-

sible to give an objective interpretation for probabilities defined over centred events: in what sense do facts about the world *determine* the chances with which centred events can occur? This is evident in particular when we consider centred events that are purely indexical, as in examples 1 and 2. In these cases, the only type of uncertainty present seems to be relative to the agent's own location and not to any external features of the world.

A closer look at the examples will, however, reveal that it may be possible to formulate an objective interpretation of probability for centred events. This section explores two approaches that an objective interpretation could follow. Firstly, I will look at how the notion of a relative frequency could be operationalised in the context of centred events. Secondly, I will turn to a functional characterisation of objective chance (in line with a propensity or a best-system interpretation of probability) and examine whether and how it could be translated in the context of centred events.

5.5.1 *Centred relative frequencies*

According to the frequency interpretation, the probability of an event E is the relative frequency of its occurrence within a suitable reference class. The first thing to do, therefore, in order to operationalise a frequency interpretation of probability for centred events, is to identify the relevant reference class and a method for 'counting' the sequence of occurrences of E.

To see how we might go about this task, let us go back to example 1. In that example, Ann knows that she is going to be woken up twice in the early hours of the morning by the town bell; she also knows in advance that she will go back to sleep immediately after hearing the sound and forget about it. Supposing that Ann just woke up hearing the sound of the bell, what is the probability

that it is 3am? In other words, we want to know the probability of the centred event $E^*_{3am} = \{(w, 3am)\}$, given that an awakening takes place: in this case, the relevant reference class is the class of all the possible awakenings. This class contains two possibilities, $(w, 3am)$ and $(w, 4am)$, corresponding to the two time points at which Ann could wake up according to the story. Having identified the appropriate reference class is still not enough to determine the objective probability of the centred event E^*_{3am} in which we are interested. For this, we also need to have additional information about the relative frequency with which E^*_{3am} is produced. In other words, we need to know the proportion of E^*_{3am}-occurrences within the reference class.

We could think of empirical ways to gather this data. Supposing that the experiment is conducted many times (in the story, the town bell follows the same pattern every day), we could record each time that Ann wakes up, noting next to each awakening the time at which it happens. Comparing the number of 3am awakenings to that of 4am awakenings, we would see that they are almost exactly equal.[43] This provides the second ingredient to identify the objective probability of E_{4am} as its relative frequency. Since E^*_{3am} occurs (almost exactly) half of the times within the relevant reference class, the objective probability assigned to E^*_{3am} by the frequency interpretation is $1/2$.

Example 1 only contains indexical uncertainty. When only uncentred uncertainty is present, the centred frequentist reduces to the 'ordinary' case. To see how it would deal with a mix of uncentred and indexical uncertainty, we can use example 3. Here Ann is uncertain both about the time, and whether the bell rings once or twice. Using the same method to register the occurrences, every time that Ann is woken up by the sound we can note down the awakening, accompanied by the time and heads/tails. If the experiment were re-

43 The qualification 'almost exactly' is necessary to account for the fact that the number of occurrences would not be equal if the total number of awakenings that have been registered is odd.

peated many times, we can expect the total number of 3am awakenings to be twice that of 4am awakenings. Moreover, the total number of Tails awakenings would also be double that of Heads awakenings. On average, therefore, whenever Ann wakes up the frequency interpretation assigns a probability of 1/3 to it being 4am, and equally a probability of 1/3 to Heads.[44]

The arguments just offered are based on an identification of relative frequencies with actual sequences of occurrence of a certain event. As discussed in section 3, this may be problematic when E is a one-off event that has never taken place or been observed before. For instance, we could imagine that the town Bell is only due to ring in the specified pattern on a particular day and we want to determine the objective probability of it being 4am, or of the coin toss coming up heads, when Ann will wake up. If this is the case, then actual frequencies seem to be a poor guide to determine objective chances.

As in the case of 'ordinary' relative frequencies, the problem can be solved by using hypothetical limiting frequencies instead of actual ones. Examples 1 and 3 involve centred uncertainty between different points in time. A feature of time is that it is linearly ordered. It is this feature of time that enables us to appeal to hypothetical frequencies to make a definite prediction.[45]

Other dimensions over which centred uncertainty could be present, however, do not have the same structure. Space, for instance, does not typically have a linear order. If the centred uncertainty involves spatial locations, the fact that an uncentred event occurs does not imply that any particular location within it is reached, or that it can be reached only once. We can illustrate this point using example 2. In that example, Tom is uncertain about which of two points

44 Note that 'heads' in this context may not be counted as an uncentred event. By the definition, an event E is uncentred if, for all $c_i, c_j \in C$ whenever $(w, c_i) \in E$, $(w, c_j) \in E$. But the event 'heads' does not contain all the possible centres; in particular, it does not contain the centre corresponding to the 4am time point.

45 See also Chapter 6, §6.4.3 and Chapter 7, in particular §7.2, §7.3.1 and §7.4.

on the map, x or y, corresponds to his actual position, and we want to assign a probability to the centred possibility that he is at x (that is, the centred event $E^*_x = \{(w, x)\}$). The relevant reference class in this case is that of all instances where Tom's position could be either x or y.

As for example 1, we can imagine to record Tom's position whenever he is in the relevant circumstances, repeating the experiment many times. Unlike example 1, however, here there could be no prior expectation that the recorded occurrences would follow a specific pattern. In the absence of additional information about how Tom could reach either location, neither is guaranteed to ever be reached or to be reached a certain number of times. The value of the probability assigned to E^*_x by the frequency interpretation would therefore vary in accordance with the actual sequence of occurrences of E^*_x that would be recorded. In the case of example 1, the linear order of time ensured that each location within the actual possible world would be reached exactly once. This gives us specific information regarding the process that generates the sequence of occurrences of the relevant centred event, using which is possible (in principle, if it exists) to determine the relative frequency of the event. But in example 2 there is no specified process that generates the sequence of occurrences of spatial locations. Because this information is lacking, we can't define a limiting frequency for the centred event in this case.

As this discussion shows, the frequentist interpretation of probability is in principle applicable to centred worlds. With respect to the criteria of adequacy for interpretations of probability given in §5.2.1, the centred frequency interpretation inherits the issues of the ordinary version. A special issue arises when the set of possible locations does not possess a certain structure (for example, if they are not linearly ordered). When this is the case, the hypothetical relative frequency of a centred event appears to be underdetermined.

5.5.2 *Functional characterisation of centred objective chance*

Other versions of the Objective interpretation of probability identify objective chance with the propensity of a certain physical object or setup to produce a given outcome. But the propensity interpretation of probability seems ill-suited to capture the particular nature of uncertainty involved in centred possibilities. The main worry here is that the kind of uncertainty related to centred possibilities generally is, by definition, not (or not just) uncertainty about the world; but it is only the latter that may be legitimately viewed as *objective*. This worry builds on the intuition that objective chance should be viewed specifically as a property of physical systems, and that the latter may not involve centred possibilities. I will call this the 'metaphysical' view, because it essentially views objective chance as a property that pertains to some specified objects.

On what I have called the functionalist view, however, objective chances are characterised by the functional role that they play. On the functionalist view, we can be more liberal about what kind of objects could bear objective chances. In principle, any probability function can count as an objective chance function just in case it satisfies the appropriate conditions. Building on Shaffer (2007) and Glynn (2010), List and Pivato (2015) put forward six characteristic properties for objective chance. In their paper, they present a framework where probabilities are defined over ordinary possible worlds, but that formulation can be extended to centred worlds. To do this, I will need to introduce a new notion, that of a context, which will enable us to translate List and Pivato's framework for the purposes of this chapter.

A context $k_1, ..., k_n \in K$ represents the specific standpoint from which the probability of a centred event E^* is evaluated.[46] This standpoint will include, for in-

46 For simplicity, I consider a finite K, but the discussion could be extended to the case where K is infinite.

stance, the evidence that is available to the agent or a description of the state of the world. Formally, each k_i is defined as the smallest centred event that contains all the centred worlds that are not currently ruled out by the evidence. The set K of all contexts is therefore a subset of the algebra of centred events S^*.[47]

I can now state the six properties of objective chance. The wording of the definitions that I am going to present is closely adapted from List and Pivato. Properties 5 to 9 were originally brought up and discussed by Shaffer (2007), while property 10 is originally due to Glynn (2010).

The first characteristic property of objective chance is a version of Lewis' Principal Principle (see §5.2.4, def. 14). It serves to establish a link between objective chances and subjective probabilities:

Property 5 (Chance-credence)**.** If an agent, in context k, were to receive the information that the objective chance of some centred event $E^* \subseteq \Omega^*$ is ch, he or she would assign a degree of belief of $cr = ch$ to E^*, no matter what other admissible information he or she has.[48]

Property 5 only requires that the agent sets their subjective credence function equal to the chances when the latter are known to her. This hypothetical

47 List and Pivato (2015) present a framework that models a system evolving over time. Time is represented by a set of linearly ordered points $t \in T$ and a state of the system is represented by $s \in S$, where S is the set of all the possible states of the system. A history h is a function that, for each time point $t \in T$, specifies a state s in which the system is at time t. In that framework, histories correspond to ordinary or centred possible worlds, and events E correspond to a collection of histories. A *truncated history* h_t is a complete history h up to time t. If the system is indeterministic, h_t may have more than one possible continuation, and h_t corresponds to the event containing all the complete histories that coincide up to h_t. For present purposes, the contexts $k \in K$ correspond to a truncated histories in List and Pivato's framework. Specifically, the truncated history h_t corresponds to the context k defined as the set of all nomologically possible continuations of history h at time t.

48 *Admissible* information is just information which bears directly on the objective chance of E^*. *Inadmissible* information, instead, is information that does not bear on the objective chance, but which could nevertheless influence the beliefs of the agent: for example, information about the future provided by an oracle, intuition, etc. See Lewis (1980).

requirement seems in principle applicable to examples 1-4 that I have used to motivate the recourse to centred worlds, as it is not immediately obvious whether objective chance is present or known to the agent.

The second and third conditions specify when a centred event E^* can have a non-degenerate objective chance.

The Chance-possibility property says that only centred events that are 'live possibilities' within the given context can have a positive chance.

Property 6 (Chance-possibility)**.** A necessary condition for a centred event $E^* \subseteq \Omega^*$ to have non-zero objective chance in a given context k is that E^* is *possible* in k.

Formally, the notion of possibility that I will adopt here is the following:

Definition 16 (Possibility)**.** An event E^* is *possible* in context k if and only if $E^* \cap k \neq \emptyset$.

In the context of example 3, for instance, both '3am' and '4am' are possible events. In example 4, 'map 1' and 'map 2' are possible events according to definition 16.

The third property, which I will call Chance-contingency, asserts that only centred events that are contingent in a given context (or, in other words, that are not yet 'settled' in that context) can have non-degenerate objective chance (that is, an objective chance that is strictly between 0 and 1):

Property 7 (Chance-contingency)**.** A necessary condition for a centred event $E^* \subseteq \Omega$ to have non-degenerate objective chance in context k is that E^* is contingent in k.

To make this condition precise, more needs to be said about what it means for an event to be *contingent* with respect to a context. Intuitively, an event is contingent when both it and its complement are possible. The definition of contingency that I will adopt here to capture this idea is the following:

Definition 17 (Contingency)**.** An event E^* is *contingent* in context k if and only if $E^* \cap k$ is a proper subset of k.

Note that a subset B of another set A is *proper* whenever $B \neq \emptyset$ and $A/B \neq \emptyset$ (in other words, B is neither empty nor coincides with A). For contingent events, this implies that both E^* and its complement overlap with the context k. The definition of contingency just given entails that Chance-contingency is a consequence of the Chance-possibility property, as can be easily checked. In example 3, the centred event corresponding to 'waking up' is not contingent, because it coincides with the context. The centred event corresponding to '3am' is contingent, because there is another distinct centred event ('4am') that is also possible in that context.

The next property, Chance-intrinsicness, expresses a regularity requirement for objective chances:

Property 8 (Chance-intrinsicness)**.** For any contexts k, k' and any centred events $E, E' \subseteq \Omega^*$, if the pair (E,k) is an "intrinsic duplicate" of the pair (E',k'), the objective chance of E in k is the same as that of E' in k'.

While the basic idea is intuitive, to make this condition precise, again it will be necessary to say more about the notion of an 'intrinsic duplicate'. In example 1, when Ann wakes up hearing the town bell there are two centred possibilities that taken together constitute the context, namely $k = \{(w,3am),(w,4am)\}$. We can imagine that the same scenario is repeated many times, as the town

bell and Ann's consequent brief awakenings take place every morning (as we envisaged in §5.2.3, discussing the frequentist interpretation). Every day sees a repetition of the same scenario; each successive repetition is obviously distinct, and yet they all intuitively share the same essential features or intrinsic properties. To capture this intuition, I will say that each successive repetition of the same scenario is an intrinsic duplicate of all the others. The Chance-intrinsicness condition therefore requires that if $ch(E) = x$ in some repetition, then $ch(E)$ should also equal x in all other repetitions.

The next property, Chance-lawfulness, requires that the objective chances of centred event should be derivable from a set of laws that regulates the behaviour of the set of all centred possibilities Ω^*:

Property 9 (Chance-lawfulness)**.** There is a set of laws at the level of Ω^* that determines the chance structure on Ω^*.

The Chance-lawfulness property expresses the basic idea that for something to count as objective, it should be possible to understand it in terms of laws. While this property can be *formulated* for centred possibilities, it directly commits us to the existence of a set of laws at the level of Ω^*. However, it is not very clear what these laws would be.

Finally, the last property of objective chance discussed by List and Pivato links objective chance with causal relevance:

Property 10 (Chance-causation)**.** If, in context k, some centred event C is positively causally relevant to another centred event E, then (except in a case of redundant causation) the chance of E, conditional on C, is greater than the unconditional chance of E.

Defining the notion of causal relevance is difficult, and it is is not immediately clear to me how this might be done for centred events. However, it should be noted that the Chance-causation property expresses a conditional requirement, which would be vacuously satisfied if the notion of causal relevance did not apply to centred events at all.

If properties 5-10 can be satisfied at the level of centred worlds, the objective interpretation of probability (in the functionalist version of the Best-system interpretation) would be applicable to centred events. In light of this discussion, I see no particular reason why a centred probability function could not in principle satisfy the six properties identified by List and Pivato (2015), and thereby play the functional role of an objective chance function. Of course, establishing this would require some more work, as some of the properties discussed – specifically, Chance-lawfulness (9) and Chance-causation (10) – would require more justification. However, as will become clearer in the next two chapters, I will not pursue this line of research further in this thesis. Instead, I will turn to focus primarily on the issues that centred worlds bring up for the subjective interpretation of probability.

Importantly, this version of the objective interpretation would satisfy the admissibility, ascertainability and applicability criteria of adequacy,[49] and the discussion in this section shows that the objective interpretation of centred probabilities would be formally viable. Interpretationally, however, this raises an interesting philosophical point. In what I have said so far, it is implicit that a context is an *epistemic* construct, while a truncated history h_t – at least, in the interpretation of List and Pivato's framework – is a *ontic* notion. So, the upshot of this discussion is that if the account of centred chance that I have outlined

49 In order to prove that it would also satisfy the compatibility criterion, one would need to show that, supposing that there are (i) a chance function at the level of ordinary possible worlds that satisfies the six properties, and (ii) a chance function at the level of centred worlds that also satisfies the six properties, then the two chance functions must stand in the relationship required by definition 15.

in this section is successful, it will give rise to a notion of objective chance as relative to an epistemic standpoint.

5.6 SUBJECTIVE INTERPRETATION

The subjective interpretations of probability identifies the probability of a centred event E with the degree of belief $cr_i(E)$ that agent i assigns to E. In Section 5.2.4, we have seen that the subjective interpretation has different versions, depending on whether cr is taken to represent the beliefs of an actual agent, or whether some further rationality constraints are imposed. In this section, I am going to focus specifically on the Bayesian version of the subjective interpretations of probability, setting the unconstrained subjectivist version aside.

The two core theses of Bayesianism are *probabilism* and *conditionalisation* (see §5.2.4). While probabilism represents a synchronic requirement on cr, conditionalisation sets a dynamic constraint on how cr may change over time as new information is learned by the agent. In this section, I consider how the Bayesian interpretation of probability can be formulated and defended in the context of centred worlds.

5.6.1 *Probabilism and centred events*

As long as the set of possible centred events S^* forms an algebra on Ω^*, it is possible to define a probability function p that assigns a probability value to every event in S^* (see §5.1). The formal machinery that is needed to define probability functions in the case of centred worlds is exactly the same as in the case of ordinary possible worlds. Therefore probabilism, which is the thesis that an

agent's rational credences cr should conform to the probability calculus, can be formulated in the case of centred worlds.

As a descriptive thesis, probabilism is implausible: we have seen in §5.2.4 that the credences entertained by actual agents often do not conform to the axioms of probability. However, the justifications for probabilism offered by Bayesians generally focus on the normative aspects.

Dutch Book arguments and accuracy arguments are used to show that if an agent does not distribute her credences among the possible events in accordance with the probability calculus, she will be liable to behave irrationally and her credences will be less accurate. Representation theorems also show that, whenever an agent has consistent preferences (a mark of practical rationality), her beliefs and desires can be represented by a pair of a probability and a utility function.

All these arguments do not depend on the specific content of the set Ω of basic possibilities, but only on the formal structure of cr and the role it plays for the practical rationality of the agent. Therefore, they carry over to Ω^*. For the remainder of this section, I will assume that probabilism (the synchronic thesis) holds, and will instead concentrate on the updating rule (the diachronic thesis of the Bayesian interpretation).

5.6.2 *Centred conditionalisation*

The rule for updating credences that is generally advocated by Bayesians is conditionalisation. As we saw in §5.2.4, this provides a diachronic constraint on cr, specifying how credences should change over time as new information is learned by the agent. The arguments for conditionalisation reviewed in §5.2.4

assume that the content of credences are ordinary events, that is sets of possible worlds. I now turn to consider whether conditionalisation is also justifiable as a rule for updating credences over centred events.

A first thing to note is that the arguments used to justify conditionalisation in the ordinary case should also be automatically applicable to centred events, when we restrict the possibilities to the class of *uncentred* events (see definition 7). This is because, by Correspondence (definition 6), these events correspond to ordinary events. But when the set of possibilities is not restricted to uncentred events, conditionalisation seems harder to justify and the arguments used in the ordinary case give problematic results. To illustrate, let us go back to our two motivating examples. In example 1, before going to sleep Ann anticipates that she will be woken up twice by the town bell, at 3am and at 4am. Let t_1 denote the time before going to sleep, and t_2 the time when Ann first wakes up during the night. At t_1, Ann is certain that it is not 3am, so she assigns a credence $cr_{t_1}((w,3am)) = 0$. Upon waking up at t_2, Ann does not know what time it is, but her credence that it is 3am is now greater than zero, as she now entertains this as a live possibility, so $cr_{t_2}((w,3am)) > 0$. In example 2, analogously, before going out on his walk, at time t_1, Tom assigns credence $cr_{t_1}((w,x)) = 0$ to the possibility that he is at location x. As he checks the map at t_2, however, his credence that he is at (w,x) is greater than 0. The change in Ann's and Tom's credences could not have been brought about by conditionalising on any new piece of evidence. According to definition 9, conditionalisation is not applicable if the prior degree of belief in an event is zero (see §5.2.4).

It is important to note that even in the case of centred possibilities, conditionalisation can go through when the centre remains fixed. However, the examples above serve to show that, as a diachronic rule for updating beliefs, conditionalisation is systematically violated when applied to centred possibilities in all

those cases where the centre is not fixed, but shifts as the agent engages in belief revision. While the diagnosis of the problem is relatively uncontroversial, the solutions that have been offered in the literature on centred conditionalisation vary greatly and no clear consensus seems to emerge relative to what would be the best strategy to address the issue. On the one hand, it could be argued that conditionalisation should be rejected entirely. This line also receives support by some arguments to the extent that conditionalisation could be problematic even in the ordinary case (see Titelbaum (2016a). I will have more to say on this point in Chapter 6).

On the other hand, some may see it as simply a problem for Bayesians to find an alternative rule that should apply when the centre is subject to shift. Many different rules have been proposed in the literature, such as 'compartmentalised', 'generalised' and 'restricted' conditionalisation. Most of these proposed rules tend to isolate the cases where ordinary conditionalisation seems to go through, and propose ways of extending the rule to the other cases on a more or less ad-hoc basis (Titelbaum, 2016b). I will examine some of these proposals in more detail in the next chapter, and propose a solution to the puzzle.

In conclusion, the problems posed by centred conditionalisation raise serious issues with respect to the applicability of the Bayesian version of the subjective interpretation to centred events, when the centre is not fixed but subject to change as the agent engages in belief revision.

5.7 DISCUSSION

This chapter has considered which interpretations of probability are available for centred events. In §5.2, I have reviewed the main interpretations of probability that have been put forward in the literature for ordinary events, which

are grouped into three main categories: Logical, Objective and Subjective. In section 5.3, I have formulated the same interpretations of probability for centred events. As emerged from the discussion, some version of each of the three main categories of interpretation is applicable to centred events.

It is interesting, in particular, that an objective interpretation of probability can be formulated for centred events. Since they were introduced by David Lewis, centred worlds have been used in the context of subjective probabilities and this seems natural, since they are taken to represent a type of uncertainty that is linked to the position of an agent or subject. However, the discussion in §5.5 shows that it is possible, under some circumstances, to identify the relative frequency of centred events, when the mechanism by which they are generated is known. Even if one does not accept the frequentist interpretation and does not identify probabilities with frequencies, a definite relative frequency can still be taken to indicate the underlying objective chance of an event.

The functionalist characterisation of objective chance can also be applied to centred events, as outlined in §5.5. This means that objective chances could be identified in the context of centred events, although some interpretational issues remain open: in particular, it would be useful to find a working definition of causal relevance in the context of centred events. Moreover, the functional account of objective chance commits us to the existence of laws at the level of centred worlds, but it is unclear what this new set of laws would be like.

For what concerns the subjective interpretation of probability, the main issue that emerged relates with its application to centred events is conditionalisation. The two central theses of Bayesianism, probabilism and conditionalisation, can both be formulated for centred events, as discussed in §5.6. However, only probabilism seems to transfer straightforwardly from the ordinary case. The problem with formulating conditionalisation in the centred context is that

this rule for updating credences seems to break down when applied to certain types of centred content. As discussed in section §5.6.2, in the finer-grained context of centred worlds conditionalisation only seems applicable to *uncentred* events or whenever the centre remains fixed as the agent receives information and engages in belief revision. In all other cases, this updating rule does not appear to be justifiable as it does not capture how credences ought to change with respect to a shifting centre. Bayesians have responded to this difficulty by devising new and different diachronic rules, but there is currently little consensus as to what is the best strategy in this context. Self-locating uncertainty therefore appears to raise a puzzle for Bayesian accounts of probability, which will be the focus of discussion in the next chapter.

6

A DIACHRONIC PUZZLE

Self-locating or centred credences are generally taken to cause problems for the diachronic Bayesian principle of conditionalisation. Two responses are available to Bayesians: revise conditionalisation, formulating a different updating rule that can account for self-locating credences; or give an account of self-locating credences within a framework that is compatible with conditionalisation. Numerous proposals in the literature have advocated different versions of the former response (see Titelbaum, 2016b). To the contrary, I argue for the latter, by showing how self-locating credences can be expressed within a framework that is compatible with standard Bayesian updating. I contrast my account with the one put forward by Moss (2012), arguing that Moss's account can be derived as a special case within my framework.

6.1 BAYESIAN UPDATING AND SELF-LOCATING UNCERTAINTY

In previous chapters, I argued that centred worlds can be used to represent self-locating uncertainty, and that it is possible to define probabilities over centred events. Moreover, I argued in chapter 5 that probabilities defined on cen-

tred events are compatible with any of the main interpretations of probability, including logical, objective, and subjective interpretations. In discussing the subjective (or Bayesian) interpretation, I focused only on the synchronic component of this interpretation, namely the thesis known as *probabilism*. According to probabilism, the graded beliefs or credences of a rational agent should always conform to the probability calculus. Standard arguments in favour of probabilism include Dutch Books and expected accuracy arguments, which, as we saw, carry over naturally to the case of credences defined on centred events.

While the classic presentation only presuppose probabilism as a basic commitment for the subjective Bayesian interpretation of probability (see de Finetti (1937); Savage (1954); Ramsey (1931)), more recently this has come to be supplemented by a diachronic constraint – *conditionalisation* – relating the credence functions that represent a rational agent's beliefs at different points in time. It works like this: suppose that cr_{t_0} is a credence function representing your beliefs at time t_0, and at a later time t_1 you learn a new piece of evidence E (and nothing more). Then, for any event A, your credence in A at t_1 should be equal to your credence in A at t_0, conditional on E. In other words, your credences satisfy conditionalisation if they change over time according to the following rule (see Chapter 5, §5.2.4):

Definition 13 (Conditionalisation)**.** $cr_{t_1}(A) = cr_{t_0}(A|E)$ (where E is the total evidence that is learned between t_0 and t_1).

Conditionalisation is supported by diachronic versions of Dutch Book arguments (see Lewis (2010); Skyrms (2009)) and expected accuracy (see Greaves and Wallace (2006); Easwaran (2013)). These results all tend to show that if an agent updates their beliefs over time in a way that violates conditionalisation, they will be prone to accept sequences of bets that guarantee them a sure loss (in the case of Dutch Books) or that is guaranteed to make their beliefs less ac-

curate. Williams (2012) examines the relationship between Dutch Book and accuracy arguments in a more general setting, investigating accuracy-domination arguments for conditionalisation that can be generalised to non-classical logical backgrounds.

While the results I have mentioned offer powerful arguments in favour of conditionalisation, it is generally agreed that self-location causes trouble for this principle (see Titelbaum (2016b)). To see why, consider how an agent's credences in self-locating propositions such as 'It is now Monday' or 'I am in London' are subject to change over time. Right at this moment, for example, I am certain of the self-locating proposition that it is Monday (A), and I am also certain of the self-locating proposition that it is not Sunday (not B). However, yesterday I was certain of the centred proposition B, that it was Sunday. So, I could not have arrived at my present beliefs via conditionalisation, because updating via this rule preserves certainties (so, if I was certain of B yesterday, I cannot be uncertain of this same proposition today), and makes it impossible to become certain of propositions to which one previously assigned 0 credence (so, if I assigned 0 credence to A yesterday, I cannot become certain of it today).

This creates a puzzle for Bayesians. On the one hand, conditionalisation is a widely accepted principle for updating credences, supported by powerful arguments. On the other hand, it does not seem to apply to the case of self-locating or centred propositions. So, Bayesians appear to face a difficult choice: either abandon conditionalisation, perhaps in favour of some other updating principle; or try to fit self-locating uncertainty in a framework that is compatible with conditionalisation. Titelbaum (2016b) surveys a number of attempts in the literature that take up the former option, giving rise to a varied family of proposals for centred updating schemes. The second option is the one that I will be arguing for in this chapter.

6.2 CENTRED UPDATING SCHEMES

As Michael Titelbaum notes in a recent survey article on the topic,

> [t]he current consensus in the self-locating credence literature is that obtaining a general updating scheme for degrees of belief in both centered and uncentered propositions requires us to alter (or at least supplement) conditionalisation in some way (Titelbaum, 2016b, p. 667).

This default position has given rise to an extensive literature on the topic of centred updating rules (Titelbaum himself notes that this has become almost a 'cottage industry'), which Titelbaum categorises into three families: 'shifting schemes', 'stable base schemes', and 'demonstrative schemes'. Shifting schemes work by providing a diachronic rule that links an agent's self-locating credences in indexical propositions – such as 'Today is Monday' and 'Yesterday was Sunday' – at different times. Stable base schemes, on the other hand, focus on the uncentred propositions that are part of an agent's beliefs at any point in time. These propositions form a 'stable base', in the sense that – contrary to the centred propositions an agent believes – they can be updated via conditionalisation, while the centred propositions that an agent's believes change according to different rules, which complement conditionalisation for genuinely *de se* beliefs. Finally, demonstrative schemes are built around the idea, defended by Robert Stalnaker, that 'belief about where one is in the world is always also belief about what world one is in' (Stalnaker, 2008, p. 55).[50] Proponents of demonstrative schemes seek to establish an equivalence between self-locating and non-self-locating propositions. Since non-self-locating propositions can

50 See also my discussion of Stalnaker's position in chapter 4.

be updated via conditionalisation, this allows demonstrative schemes to define a natural extension of conditionalisation to self-locating propositions.

As I mentioned above, a Bayesian can respond in two ways to the problem that self-locating uncertainty poses for the standard account of conditionalisation. Shifting schemes and stable base schemes have in common that they both constitute instances of the first type of response to this problem, as they both devise essentially new updating scheme to deal with self-locating propositions. Demonstrative schemes, on the other hand, can be seen as attempts to take the second option – trying to fit self-locating uncertainty in a framework that makes it compatible with conditionalisation. This feature makes demonstrative schemes especially relevant to my discussion, as I will also be advocating taking this second option.

Updating schemes from each of the three families, Titelbaum argues, suffer from different blind spots, making them inapplicable across a significant range of situations. Shifting schemes break down when an agent is uncertain about what is the present time, so that they cannot be used to assign probabilities to indexical claims such as 'It rained yesterday', if the agent is uncertain about the current day. Stable base schemes (including those put forward by Titelbaum (2008) and Briggs (2010)) have a different blind spot: in Titelbaum's words, they 'try to cash out an agent's credences entirely in uncentered terms, which are straightforwardly manipulable by conditionalisation. But when an agent has no qualitative way of identifying a day (or a place, or a person) [...] such schemes fall short' (Titelbaum, 2016b, p. 675). Finally, demonstrative schemes, in particular the one put forward by Stalnaker (2008), use demonstrative reference to establish a correspondence between centred and uncentred propositions. However, Titelbaum argues that they are also not successful in handling cases where the set of epistemic possibilities increases, for example as a result of becoming uncertain about the present time. When this happens, condition-

alisation alone does not determine the credences that one should assign to the uncentred propositions that only become available at the later time, as we will see in a moment.

As Titelbaum points out, the respective blind spots limit the applicability of the updating schemes in each of the three families. A particular problem, known in the literature as the *Sleeping Beauty problem*, is especially interesting in this regard, as it appears to lie within the blind spots for each of the updating schemes identified by Titelbaum. This explains the interest that the problem has generated in the literature. I will discuss this problem in detail in the next chapter.

6.3 DEMONSTRATIVE SCHEMES

As we saw in the last section, Titelbaum concludes his review of centred updating schemes by noting that any centred updating scheme in the literature suffers from a blind spot, which makes the schemes from all the three families he identifies inapplicable across a significant range of situations. The blind spot he identifies for demonstrative schemes is a result of the reliance of such schemes on direct reference. To illustrate the kind of difficulty that this poses to demonstrative schemes, he considers the following example:

> *Roger Foretold*: On July 4th Roger knows that he's about to begin an extended period of sleepings and awakenings [...]. This process begins, and some number of days later Roger finds himself awake, uncertain which awakening it is or how long he's been asleep. On this awakening Roger looks out the window, sees clouds, and becomes 0.7 confident it will rain. With which of Roger's July 4th

> credences is this 0.7 credence coordinated? (Titelbaum, 2016b, p. 676)

Titelbaum argues that the example of Roger Foretold raises a problem for the demonstrative scheme put forward by Stalnaker (2008). To see this, let us try to apply Stalnaker's account to this case. When Roger wakes up uncertain about which awakening it is or how long he's been asleep, his uncertainty about the current time corresponds to uncentred uncertainty about what the world is like. The object that he can now demonstratively refer to as *this moment*, the present moment, occurs at different points in time in different possible worlds, but could occur only once in each possible world (this is implied by Stalnaker's propositionality condition, which I discussed in Chapters 3 and 4). So, Stalnaker concludes that Roger's uncertainty about the time when he wakes up is really uncentred uncertainty about what the world is like – namely, whether the actual world is one in which *this awakening* takes place at time t. This indicates that Roger's credences when he wakes up uncertain about the time are distributed across a set S of possibilities, each corresponding to a possible world where *this particular awakening* takes place at a different time. The problem, however, is that the set S corresponding to the possible worlds that are epistemic possibilities for Roger when he wakes up is not a subset of the set of possible worlds (call it S_{-}) that were epistemic possibilities for Roger on July 4th, before going to sleep and losing track of time. This is because on July 4th, Roger knows exactly which sequence of awakenings he is going to experience. What changes between the two times (July 4th, and the moment he wakes up) is that, upon waking up, he gains the ability to refer demonstratively to an object ('this moment') to which he did not have direct referential access before.[51] So, the set of epistemic probabilities S_{-} available to Roger on July 4th is smaller than the set of epistemic possibilities S that are available to him at the time he

51 See also Weber (2015) for a related discussion.

wakes up, some days later. However, this change could not be captured by conditionalisation, which can only work to narrow down a set of possibilities – not expand it.

Note that the example of Roger Foretold poses a problem for Stalnaker's demonstrative scheme, for, as Stalnaker (2011) himself acknowledges, it will require more than conditionalisation to account for rational updating in cases like that of Roger Foretold, where agents lose track of their identity, time or spatial location. Since memory loss appears to be involved in this example (by some point, Roger forgets the number of awakenings he has already been through), one may ask whether this really poses a specific problem for Stalnaker's scheme, as opposed to presenting a case where we would naturally expect conditionalisation to break down. As I recalled in the first section of this chapter, conditionalisation relates the credence functions of an agent at two different times, under the assumption that between those times the agent does not lose information. This is something that appears to happen to Roger Foretold: between the 4th of July and the awakening under consideration, he has lost the memory of the number of his previous awakenings.

Even if the break down of conditionalisation is to be expected (due to memory loss) and is not specific to Stalnaker's framework, we might still think that it is a problem. After all, having a perfect memory may not be a requirement of rationality, but conditionalisation does not say how agents should rationally respond to evidence loss. Moreover, in one important respect this example does not presuppose memory loss, because the time when Roger wakes up may be the first awakening in the sequence, and so no awakenings need have been forgotten for Roger to become uncertain about the current time. It is important to note this, because conditionalisation can only apply to cases where no information is lost between a time t_i and a time $t_j > t_i$. So, if the example only hinged on this kind of cognitive mishap, it would not represent a genuine coun-

terexample to demonstrative schemes; it would merely exhibit a case where conditionalisation predictably breaks down.

A recent proposal, put forward by Sarah Moss (2012), outlines an updating scheme for centred propositions which she calls 'black box updating'. Titelbaum classes this updating scheme in the same family as Stalnaker's demonstrative scheme, with which Moss's black box updating has many features in common. However, Titelbaum also briefly notes that Moss's scheme is not be vulnerable to the same counterexample as Stalnaker's scheme. This is because Moss's demonstrative scheme is not identical with conditionalisation, and it can account for cases where – like in the example of Roger Foretold – the set of possible worlds that constitute epistemic possibilities increases between two successive times t_j and t_i. In the remainder of this section, I will outline the details of Moss's proposal, before moving on to present my own proposal in the next section.

6.3.1 *Moss: Updating as communication*

For Moss, updating *de se* beliefs is just like receiving information from your earlier self, and adjusting your present beliefs in accordance to that information. In other words, assuming that you should always trust your earlier self, the information that you receive from them constrains what it is rational for you to believe at present. What type of information gets communicated in this way? According to Moss, this could not be the *de se* propositions that your earlier self used to believe. This is because these *de se* propositions are not, generally, propositions that your current self can also believe: for example, if my earlier self used to believe that it was 4pm, and some time has passed since then, my present self cannot have the same *de se* belief. Instead, according to Moss, the

information that is communicated from the earlier self to the later self are *de dicto* propositions. This, on Moss's account, is made possible by the following principle:

> [PROXY] Given a *de se* proposition, there is a *de dicto* proposition such that for any centered world compatible with what you believe, that centered world is in the former proposition just in case it is in the latter. (Moss, 2012, p. 226)

So, on this account, the content that is communicated by the earlier self to the present self are the *de dicto* (that is, uncentred) propositions equivalent with what the earlier self used to believe *de se*. These *de dicto* propositions can be communicated and believed by both your earlier and your present self. In this way, Moss argues that updating can be viewed as a particular instance of communication – not between different agents, but between different stages of the same agent at different times.

Before moving on to consider Moss's proposal in more detail, however, it is worth noting that Moss's analogy between communication and updating might not quite work as she claims (see also Pagin, 2016, pp. 286-290). To see this, let q be a *de se* proposition that the speaker, X, wants to communicate to a hearer, Y. By PROXY, q is equivalent to some other *de dicto* proposition p, given what X believes with certainty (a 'background certainty' proposition r). In other words, by PROXY, q is equivalent to $p \wedge r$, where r is a background proposition that X believes with certainty. Since p is a *de dicto* proposition, X can communicate p to Y, and Y can come to believe p (if he trusts the speaker). Now, it is easy to check that the background proposition r must be *de se*, because otherwise the conjunction of p and r would have to be a *de dicto* proposition.[52] But if so, then X cannot successfully communicate a *de dicto* proposition that is

52 The conjunction of two *de dicto* propositions is a *de dicto* proposition.

equivalent to q, unless she can also contextually communicate what the background belief r is. But this is not possible on Moss's account, since *de se* propositions cannot be directly communicated.

To illustrate, suppose that Ann is on the phone with Bob, and believes the *de se* proposition that she is hungry (q). On Moss's account, Ann cannot communicate q directly to Bob, but she can communicate some *de dicto* proposition p that is equivalent to q, given what she also believes with certainty – for example that she herself is Ann (this is her background *de se* proposition r). So, if Ann says over the phone, 'I am hungry', the content of what she communicates to Bob is the *de dicto* proposition that Ann is hungry. However, the problem with this account is that Bob might not know that he is speaking to Ann – maybe he is confused about the identity of the speaker, or he just doesn't know her, or some other reason. The point is that he might not come to believe p, unless he also has access to the background information that the speaker is Ann. So, Ann cannot successfully communicate the *de dicto* proposition p to Bob, unless she can also communicate the content of her background *de se* belief r. But this takes us back to the starting point: by Moss's own lights, since r is a *de se* proposition, Ann cannot communicate it directly to Bob.

This discussion highlights a serious problem for Moss's account of communication. However, the same issue might not arise in the case of updating, which I will describe more in detail below. Even so, this casts doubt on Moss's claim that since PROXY enables (among other things) an elegant model of communication, its 'theoretical fruits [...] may be substantial enough to justify our acceptance of the claim itself' (Moss, 2012, p. 236).

6.3.2 *Black Box Updating*

Moss (2012), following Lewis (1979) and Stalnaker (2008), uses sets of centred worlds to represent an agent's belief state. Unlike Lewis, but in accordance with Stalnaker, she further accepts that agents' beliefs sets always satisfy propositionality (see definition 1, and Chapter 4, §4.3.2). Let $Bel_{(a,t)}$ be the belief set of an agent a at time t, containing all the centred worlds that constitute epistemic possibilities for a at t. $Bel_{(a,t)}$ satisfies propositionality if it has the following property:

Definition 1 (Propositionality)**.** For any pair of centred worlds $(w,c),(w,c')$ that coincide on the uncentred component w and such that $c \neq c'$, if (w,c) belongs to $Bel_{(a,t)}$, then (w,c') does not belong to $Bel_{(a,t)}$.

Propositionality rules out that an agent could be uncertain about self-locating or centred information without also being uncertain about some uncentred information. In other words, $Bel_{(a,t)}$ can contain at most one centred world for each possible world w. Moss uses this assumption to build an updating scheme that works in two steps. On the first step, some uncentred information is communicated from an earlier to a later self, and used to derive a 'black box updated' belief set $Bel_{(a,t_+)}$. Then, $Bel_{(a,t_+)}$ is updated by conditionalising on any uncentred information that is learned by the later self. In Moss's words:

> Genuine rational updating happens in two steps. First you update as if you were in a black box. Then you conditionalize your resulting credences on what you genuinely learn. (Moss, 2012)

The general idea underlying Moss's proposal is the following. At any point in time, an agent's belief set is characterised by two components, that I will call

the *centred* and the *uncentred* components. The centred component is the set S of centred propositions that are compatible with the agent's beliefs. Similarly, the uncentred component is the set P of all the uncentred propositions that the agent believes. On a first reading of Moss's proposal, the uncentred component of an agent's belief set can be updated via conditionalisation, while the centred component is revised via a different rule. As the former (updating the uncentred component of the belief set) amounts just to standard conditionalisation, it is the latter step (revising the centred component of the belief set) that sets Moss's account apart from other updating schemes. Here is how she describes the working of black box updating:

> In hypothetical black box updating, you form beliefs on the basis of information you get from your previous self. [...] Each *de se* proposition you used to believe is equivalent with some *de dicto* proposition, given what you used to believe with certainty. This sort of *de dicto* proposition is something you can currently believe. Furthermore, you currently have some *de se* beliefs about your relation to your previous self. So you can also currently believe some *de se* propositions: the consequences of your current *de se* beliefs and your old *de dicto* information. (Moss, 2012)

Based on what Moss says, we can break down black box updating into the following smaller sub steps:

1. Identify the uncentred component of your earlier self's beliefs;

2. Identify the centred component of your later self's beliefs;

3. Combining 1 and 2, construct a hypothetical belief set for your later self.

If some new information is learned by the later self, which was not initially available to the earlier self (and thus is not reflected in the earlier self's belief set), the resulting hypothetical belief set that is constructed via steps 1-3 can then be updated via standard conditionalisation.

To better evaluate Moss's proposal, I will need to say something more about how the new hypothetical belief set can be constructed in step 3 of black box updating. In many cases, this will just involve shifting the centred component of each centred world that is in your earlier belief set. For example, suppose that you go to sleep just after looking at the clock at 11pm on Sunday, and wake up hearing the alarm at 6am the following day, after an undisturbed sleep. At both points (on Sunday before sleeping, and on Monday upon waking) you are certain of the current time, and the uncentred information that is available to you has not changed between these times. So, your later belief set is in the following relation to your earlier one: for every centred world that is in your earlier belief set, the w components stays constant, but the c component is revised, to reflect that your current time location is 6am on Monday.

If that was all there was to the story, however, black box updating would be inapplicable to cases – like the example of Roger Foretold – where an agent predictably becomes uncertain of his or her won location. In Titelbaum's example that I presented earlier, Roger is initially certain that it is the 4th of July and that he will be going through a certain series of awakenings, and as we have seen this entails that the uncentred component of his belief set cannot be updated via conditionalisation to result in a larger set of uncentred possibilities upon waking up. Moss's treatment of this kind of cases defines an important aspect of her account of black box updating. Here is what she proposes:

> In black box updating, your credences are entirely determined by two elements: your previous credences in *de dicto* propositions,

> and your current conditional credences about your relation to your previous self. First your previous credences determine how much credence you give to any *de dicto* proposition. Then your conditional credences determine how you distribute that credence among all *de se* propositions entailing that *de dicto* proposition. (Moss, 2012)

In other words, we need to supplement the first reading of Moss's account in this respect: the uncentred component of an agent's belief set can, under some circumstances, be updated via something different than conditionalisation. The idea is the following: when the set of epistemic possibility increases between two times t and t', because – as in the case of Roger Foretold – the agent becomes uncertain about some self-locating information at t', the new hypothetical belief set for the agent at t' is obtained via the following steps:

1. Identify the uncentred component of your earlier self's beliefs;

2. Identify the centred component of your later self's beliefs;

3. If any uncentred proposition p that is in the uncentred component of your earlier self's beliefs is now compatible with more than one centred proposition s in your later self's beliefs, 'split' p into as many finer grained uncentred propositions $p_1, p_2, ..., p_n$ as needed;

4. Combining 1, 2, and 3, construct a finer-grained hypothetical belief set for your later self.

This four step-procedure is a fuller account of Moss's updating scheme, as it complements the previous three step account and it ensures that the resulting belief set always satisfies propositionality, and it is able to accommodate cases like Roger Foretold's where an agent becomes uncertain of some self-locating

information as a later time. A consequence of this account is that – as Moss acknowledges in a footnote to her paper – over time, an agent's uncentred beliefs are defined relative to an increasingly fine grained set of possibilities (Moss, 2012, p. 16). Moreover, in order to distribute credences across the new finer grained possibilities, the agent is credited with possessing a conditional probability distribution that assigns a credence to any new centred proposition compatible with the later self's beliefs, conditional on each uncentred proposition believed by the earlier self. Moss does not say much about these two consequences of her account. In particular, she does not say how the conditional probability distribution that is needed to derive the later self's hypothetical belief set in black box updating is defined.

6.4 CONDITIONALISATION REDUX

As we saw in the first section of this chapter, self-locating credences give rise to a puzzle for Bayesians, as they seem to conflict with a standard diachronic principle for updating beliefs, conditionalisation. Two possible responses are available to Bayesians: either drop conditionalisation, and formulate some other diachronic principle to guide rational belief change in its stead; or keep conditionalisation, but explain how centred credences can be updated in a way that is consistent with it. Most current proposals in the literature on self-locating credences adopt the first strategy, proposing a variety of alternative updating schemes that are intended to replace or complement conditionalisation for self-locating credences. Demonstrative schemes – notably those put forward by Stalnaker and Moss – promise initially to take the second strategy, but as we have seen in the previous section they also need to complement conditionalisation with some other principle, in order to account for belief changes in cases where the set of epistemic possibilities is expanded (like in Roger Foretold's

case). As Titelbaum shows, the great majority of centred updating schemes in the literature have some blind spot, making them inapplicable across a range of situations. Only Moss's black box updating appears to be the notable exception, since, as we have seen, it is applicable to the cases that represent a blind spot for other demonstrative schemes. Based on the discussion so far, therefore, Moss's black box updating scheme emerges as a contender to replace conditionalisation even though, as we have seen, it relies on a mysterious account of conditional probabilities that are only accessible under some special circumstances.

In the rest of this chapter, I now turn to explore the second option that is open to Bayesians confronted with the puzzle of self-locating credences. My main aim will be to show that it is possible to represent self-locating credences within a standard Bayesian framework, and that when correctly represented they are consistent with conditionalisation. Before we move on, I will need to say something more about what I take to be the correct understanding of the key terms in the debate on self-locating uncertainty.

6.4.1 *Diachronic coherence*

First of all, let us recall our initial definition of conditionalisation from §6.1:

Definition 13 (Conditionalisation)**.** $cr_{t_1}(A) = cr_{t_0}(A|E)$ (where E is the total evidence that is learned between t_0 and t_1).

As we saw in §6.1, conditionalisation is a diachronic principle, which relates an agent's credence functions at different (successive) times. Along with the synchronic principle of probabilism, conditionalisation is generally taken to be a rationality principle and is supported – as recalled in §6.1 – by diachronic

Dutch Book arguments. But just what does it mean for something to be a rationality principle? There may be different ways to interpret this question and, as I will argue in this section, this makes a difference to how we should understand conditionalisation.

Let's consider the case of probabilism first, as this principle is generally considered less controversial. Probabilism says that an agent's credence function, if rational, should conform to the probability calculus. This could intuitively be interpreted in two ways. On the first reading, probabilism expresses a condition that is necessarily satisfied by a rational credence function. On the second reading, it expresses a requirement: if you want to be rational, your credences *should* conform the probability calculus. In other words, on the first reading, probabilism expresses a standard of rationality for credence functions, while on the second reading it is interpreted as a normative statement. Each reading may be appropriate to different settings, but I think it is important to notice the differences. In particular, the first reading seems the most appropriate if we want to evaluate the rationality of an agent, based on the credence function that she has at given time t. For example, suppose that your friend Ann considers it is .2 likely that Sweden will win the world football championship in 2018, and also considers that it is .85 likely that one team between either Italy or England will win the championship. Ann's credences concerning the likelihood that each team will win the championship sum to a number greater than 1, and so do not conform to the probability calculus. Based on the first reading of probabilism, therefore, Ann's credence function is not rational, because it does not conform to the probability calculus.

Your other friend Jack, on the other hand, considers that it is .3 likely that England will win the championship, .1 likely that Italy will win, and .6 likely that one out of all the other contender teams will. Jack's credences sum to 1, and are consistent with the probability calculus. So, based on the first reading of

probabilism, you can conclude that Jack's credences (at least for what concerns football matches) are rational. Importantly, you can evaluate Jack's credence function, and say that it is rational, even if Jack himself is not aware that his credences conform to the probability calculus. Rationality is a property of his credence function, which does not require self-awareness on the part of the agent who is the subject of evaluation.

Similarly, in the case of conditionalisation, there are two salient readings of this principle. For example, suppose that at time t_0 Mary thinks it is .6 likely that England will win the world championship, if Germany is eliminated in the second round. Then, at time t_1, it so happens that Germany is eliminated and she comes to learn this and nothing else. Based on one reading of the principle of conditionalisation, her credence in England winning the championship *should* be .6 at t_1, if she wants to be diachronically rational. Based on a different reading, on the other hand, we can simply say that Mary's credence functions at t_0 and t_1 are diachronically rational if her credence in the possibility that England wins the championship at t_1 is equal to .6 (i.e. if it is equal to the conditional probability she used to assign to this possibility at t_0, conditional on the information that Germany is eliminated, which she has learned by t_1). As in the case of probabilism, this second reading of the principle of conditionalisation is more appropriate whenever our aim is that of evaluating the rationality of Mary's updating behaviour, and it does not presuppose that Mary herself should be aware of updating her credences in accordance with the principle of conditionalisation.

Much of the literature on self-locating credences appears to take the first, normative reading of conditionalisation for granted. As Titelbaum writes:

> The role of an updating scheme is to coordinate credences assigned at different times. Conditionalisation, for instance, *requires an agent*

> *to line up her unconditional credences* at a later time with particular conditional credences assigned earlier on. (Titelbaum, 2016b, p. 668, emphasis added)

This interpretive choice puts the focus on explaining how agents can identify and coordinate the content of their own credences at different times. Moss's black box updating, for example, approaches this issue by treating updating as a form of communication, where your earlier self passes on some information to your later self.

In what follows, I will adopt the second reading of the principle of conditionalisation. I think there are several good reasons to adopt this interpretation of the principle – not least, this interpretive choice also sidesteps some issues that arise if we interpret conditionalisation as a positive normative requirement.[53] A consequence of adopting this interpretation of the principle of conditionalisation, is that it eliminates the issue of explaining how an agent should coordinate credences at different times. So, the problem I am considering is not: how should I coordinate my present credences to my past and future ones? But rather: given an agent's credence functions at different times, under what conditions can we say that the agent is diachronically coherent/rational?

6.4.2 *Evidence*

As recalled in the opening section of this chapter, self-locating credences are generally taken to present a problem for conditionalisation. To see why, as Titelbaum writes:

53 Timothy Williamson defends an interpretation of the principle of conditionalisation that is similar to the one I have outlined. See especially Williamson (2000), Chapter 10.

> [...] suppose we have an agent who is currently certain that it is Tuesday. Intuitively, it is apparent there are some things that agent could learn as time goes on that would make it rational for her to decrease her certainty in that proposition. (Titelbaum, 2016b, p. 667)

While the literature on self-locating propositions often focuses on temporally centred propositions (like 'it is Tuesday' in the quote from Titelbaum's review article), the problem is actually much more general. Suppose that you are certain that it is raining, and that this is some coffee. Intuitively, there are things you could learn that would make these certainties, too, decrease. For instance, you could look out of the window and notice that it isn't raining anymore. You could finish your cup of coffee, or put it down and focus on writing – and then you may become certain that this in front of you is a computer screen.

The two examples I just gave indicate a more general pattern. Everything that you perceive around you, all the data you constantly process from your environment, appear to have this fleeting character: *this object*, *this moment*, *this place*, your own perception of yourself are constantly changing. Clearly, you do not get these data via conditionalisation: they simply appear to you in a certain way, and these appearances are not fixed, they change as you move between different places and different times. If conditionalisation breaks down in all these cases, then it looks like this should be a big problem for Bayesian reasoning, as it will make it impossible to apply the standard rationality principles even to the most familiar cases.

The extent of the problem also indicates where a possible solution lies. To see how, take a standard textbook example, repeated tosses of a fair coin. Each time the coin is tossed, you are allowed to observe the outcome ('this is a Heads' – or 'this is a Tails'). As we saw, this observation is something that you

get from your environment. You become certain of 'this is a Heads' by observing a toss of the coin, and may lose the certainty that 'this is a Heads' – for example, you may observe another toss and come to be certain that 'this is a Tails'. The information that you can extract from each observation, however, can be represented and stored in different forms, e.g. as the result of a certain toss (say, the first you observe) being a Heads. You can attach an index to each observation, so the information that you get is something that you can become certain of even after your circumstances change. In other words, by assigning an index to each observation, it becomes part of your evidence.

This way of distinguishing between observations (that have a fleeting nature, and can be expressed using indexical or demonstrative sentences such as 'this is a Heads' or 'now it is raining') and evidence, or informational content, seems very natural in the case of the coin tossing experiment. I will now argue that this can be generalised to encompass the kind of cases that have been considered problematic in the literature on self-locating credences. Then, in the next chapter, I will apply this view to the Sleeping Beauty problem, which has long been considered a test case for any account of self-locating credences, and show that my view leads to a natural resolution of that problem.

6.4.3 *A unified proposal*

We now have all the elements to present a unified framework for centred and uncentred uncertainty. Let $(\Omega, \mathscr{F}, P)$ be a probability space, where Ω is the set of all centred worlds, $\mathscr{F}$ is a σ-algebra on Ω, and P is a probability function defined on $\mathscr{F}$. We do not need to impose any requirement on $\mathscr{F}$ at this stage, so we can leave it open which algebra we take the probability function P to be defined over. For example, $\mathscr{F}$ could be the finest σ-algebra defined on Ω,

which includes all the centred propositions that are subsets of Ω. But this is not necessarily required, and depending on the application it may be more natural to take $\mathscr{F}$ to be a coarser σ-algebra on Ω. For example, if we are exclusively interested in reasoning about uncentred propositions, we might take $\mathscr{F}$ to be the σ-algebra generated by the uncentred propositions that are subsets of Ω (see chapter 5 for a more detailed discussion).

Assuming that probabilism is satisfied, an agent's credences in the propositions in $\mathscr{F}$ correspond to a probability function defined on $\mathscr{F}$. In addition, we can represent an agent's current evidence as a set of centred worlds Bel, that is measurable with respect to $\mathscr{F}$, and which contains all the centred worlds that are live possibilities for the agent. Note that unless we impose some further constraints, Bel does not need to satisfy Propositionality, but may contain pairs of centred worlds that coincide on the uncentred component, while they differ with respect to the centre. In the special case where $\mathscr{F}$ is the algebra of uncentred propositions, in particular, Bel will not in general satisfy propositionality, since it will coincide with the strongest uncentred proposition that is believed by the agent – and uncentred propositions, as we have seen, contain all the centred worlds that coincide with respect to some uncentred components.

The belief set Bel changes over time, as the agent gains (or, sometimes, loses) evidence, and as it comes to occupy a different centre. The way in which these things happen is determined by external parameters, that lie outside the scope of the framework. For example, suppose that we are considering the probability that a coin toss will land Heads. The framework allows us to model the possibilities that different outcomes will occur. However, which outcome actually occurs, and is observed (becoming part of the evidence) is not determined by the framework, but by the world. Similarly, suppose that at some point in time it is Wednesday, and at some other point it is Thursday, and at neither point I

am uncertain about my current time location. The evidence that I have at the two points in time (that it is Wednesday, or that it is Thursday) is determined by the world, and not by the framework. In other words, the belief state Bel corresponding to the agent's evidential state is exogenously determined by the world, but our framework allows us to model how a rational agent distributes his credences among the live possibilities relative to the belief state that he is in.

Let Bel_i and Bel_j be two distinct belief states, corresponding to two – possibly consecutive – evidential states of the agent relative to a given algebra $\mathscr{F}$. Since they are both subsets of Ω, Bel_i and Bel_j will be in one of the following relations between each other: either Bel_i and Bel_j are disjoint (that is $Bel_i \cap Bel_j = \emptyset$), or they have some nonempty intersection. If the latter is the case we can distinguish two special further cases: i) $Bel_i \subseteq Bel_j$, and ii) $Bel_j \subseteq Bel_i$. As it is usually defined (see definition 13 in §6.1), conditionalisation only applies to the circumstances corresponding to case ii) when the two belief states that we are comparing are not disjoint, and the later belief set is a subset of the earlier one. This is because conditionalisation is a conditional claim that relates the credence functions of an agent at successive points in time, conditional on the fact that some new evidence is learned (and none lost or revised) between those times. In all other cases except ii), this condition is not met: whenever Bel_j is not a subset of Bel_i, there must be at least one centred world that is not contained in Bel_i (because it is ruled out by that evidential state), but which is contained in Bel_j (that is, it is considered as a live possibility in that state).

The discussion so far indicates that the applicability of conditionalisation, as it is ordinarily understood, is severely limited when the probability space is defined on the set of centred worlds, as the successive belief states of an agent often correspond to disjoint sets of centred worlds. This appears to be an issue,

especially if one adopts the normative reading of the principle of conditionalisation that I considered in §6.4.1. If the aim is to give a normative principle that an agent could explicitly follow to update his credence function over time, then conditionalisation would remain silent in all the cases where the evidence available to the agent changes in a way that is not compatible with the condition that the later belief set be a subset of the earlier one. The issue is not confined to centred uncertainty (uncentred evidence can also change in ways that do not satisfy the condition for the applicability of conditionalisation), but the violations appear to be much more widespread. Moreover, while in the case of uncentred uncertainty the changes that are incompatible with conditionalisation can be described as straightforward cases of evidence loss – and thus dismissed as involving a form of irrationality – it is much more difficult to dismiss the changes in centred belief sets as irrational. The fact that an agent can occupy different centres at different times seems just a fact of the world, and is much more difficult to dismiss as a simple failure of rationality. This is at least one reason motivating the recent interest in the literature on self-locating uncertainty.

As I explained in the introduction to this chapter, mostly the attempts in the literature have focused on finding some extension of the principle of conditionalisation that may be applicable to the cases that are not covered by the original principle. However, drawing on the alternative reading of the principle of conditionalisation that I described in §6.4.1, I would like to propose a different solution to the puzzle, that I think vindicates the intuitive appeal of the principle.

First, let us recall the standard formulation of conditionalisation and consider more closely the cases in which it is applicable. Here is the definition again:

Definition 13 (Conditionalisation)**.** $cr_{t_1}(A) = cr_{t_0}(A|E)$ (where E is the total evidence that is learned between t_0 and t_1).

As we have seen, the condition that the only difference between the agent's belief sets at t_0 and t_1 is that the evidence E (and nothing more) is learned by time t_1 is not always satisfied in practice. In all these cases, the change in the agent's credence function between times t_0 and t_1 does not respect conditionalisation, and so – based on the reading of the principle that I defended in §6.4.1 – the agent's credence functions are not diachronically rational. This verdict, however, may seem a bit harsh. After all, as we have seen, especially when an agent's evidential states correspond to sets of centred worlds, there seems to be nothing especially irrational to the way the agent's evidence changes, as this is determined by processes that are not under the agent's control. As long as the agent responds in a coherent way to different pieces of evidence, we should not judge his credence function as irrational. The issue is, relative to what standard can we assess diachronic coherence?

While the evidence that is available at any time changes in a way that is externally determined, there is a feature of the agent's credal state that appears to remain constant. This is the probability space $(\Omega, \mathscr{F}, P)$ that we introduced to model the agent's beliefs, where P corresponds to the unconditional probability that the agent assigns to the events in the algebra $\mathscr{F}$, before any evidence is factored in. In other words, P corresponds the prior credence function of the agent, relative to Ω. Since – by definition – all the possible belief sets are subsets of the sample space Ω, this ensures that the condition for the applicability of conditionalisation is always satisfied when we take cr_{t_0} to be equal to the unconditional probability function P. So, as long as the agent's prior unconditional probability function remains constant, we can always take P to play the role of cr_{t_0}, and use the principle of conditionalisation to evaluate whether

the agent's credence function cr_{t_1}, relative to the evidence available at t_1, is diachronically coherent.

We can use the case of Roger Foretold to illustrate how this proposal will work in practice. In that example, Roger knows that he will undergo a given series of awakenings over a few days. While he knows all of this exactly prior to being put to sleep, upon waking up he is uncertain about which day it is. To keep things simple, we can model Roger's case as involving only self-locating uncertainty. Suppose that the day he is initially awake is Sunday, and then the sequence of awakenings that he knows he will undergo will take place on Monday and on Tuesday. We can represent Roger's centred possibilities by setting the sample space $\Omega = \{(w,s),(w,m),(w,t)\}$, $\mathscr{F}$ as the finest algebra on Ω, and P as Roger's prior unconditional probability function on $\mathscr{F}$. On Sunday, Roger's belief set is $Bel_i = \{(w,s)\}$, while upon waking up (not knowing what day it is) his belief set is a different $Bel_j = \{(w,m),(w,t)\}$. Since Bel_i and Bel_j are disjoint, conditionalisation does not directly apply. But both Bel_i and Bel_j can be evaluated relative to P. This means that if he is diachronically rational in the sense we have discussed, Roger's credence in the centred proposition $\{(w,m)\}$ upon waking up will be equal to the prior probability of $\{(w,m)\}$, conditional on Bel_j, or $cr_j(\{(w,m)\}) = P(\{(w,m)\}|\{(w,m),(w,t)\})$.

Roger's case illustrates a more general feature of my account. The key point here is that we can evaluate the diachronic rationality of an individual agent, provided that the agent has a prior probability function P, that is defined for the whole of Ω, and which does not change between different points in time. If the agent is rational, then the probability function that corresponds to her credences at any given point will be fixed by the prior probability function P together with the current evidence available to the agent.

6.4.4 *Possible extensions*

As we have seen, conditionalisation is applicable whenever the agent's belief set at t_1 is a subset of his belief set at t_0. Importantly, this does not require that the agent's belief sets satisfy Stalnaker's condition of Propositionality. This is I think a good feature of my account, as it makes it applicable without modifications to centred and uncentred propositions. As we saw in chapter 4, uncentred propositions partition the set of centred worlds Ω into indifference classes of centred worlds that coincide with respect to the uncentred component. Therefore, by definition, belief sets that contain uncentred propositions will not in general satisfy propositionality.

Even for belief sets that do satisfy propositionality, my account is not guaranteed to deliver the same results as Moss's black box updating, as this will depend on the particular probability function P that is taken to correspond to the agent's prior credence function. Considering which features of P would be necessary to deliver Moss's black box updating highlights an interesting possible extension of the framework. We can recover Moss's black box updating as a special case within the framework I have outlined, by imposing two constraints: i) that all belief sets satisfy propositionality; and ii) that P assigns equal probabilities to any two centred worlds (w, c) and (w, c') that coincide with respect to the uncentred component w. Thus, this is an interesting consequence of my proposal, as it both generalises the intuitions that may be behind Moss's account, and it explicitly accounts for the technical assumptions that were left implicit in Moss's discussion (concerning the agent's underlying conditional probability distribution over centred possibilities).

6.5 CONCLUSION

In this chapter, I have considered a puzzle that self-locating uncertainty creates for Bayesian accounts of rational credences. Self-locating uncertainty has been taken to challenge the Bayesian principle of conditionalisation. As a result, different diachronic principles have been proposed in the literature, either to supplement or to replace conditionalisation. A particularly interesting proposal is that put forward by Moss (2012), which I analysed in this chapter. Moss's account of black box updating is built on the assumption of the Stalnakerian account of *de se* beliefs that I discussed in Chapters 3 and 4. The unified framework that I have presented, on the other hand, is consistent with the Lewisian account of *de se* beliefs discussed in Chapter 3, and as such it is able to provide the conceptual tools to model both modes of reasoning about *de se* beliefs that I identified in that chapter, namely the cartographer mode and the pathfinder mode.

I have argued that my account has several theoretical advantages over the competing accounts: it does not require a revision the standard Bayesian principles of rationality (under the interpretation that I have defended), it clarifies the conditions under which an agent's credence functions at different times may be considered as diachronically rational, and it gives a unified treatment to updating on both centred and uncentred evidence. In the next chapter, I will apply the framework described in this chapter to what has been taken to be a test case for any theory of self-locating uncertainty, the Sleeping Beauty problem, and show that it leads to a natural resolution of that problem.

7

BAYESIAN BEAUTY

The Sleeping Beauty problem has attracted considerable attention in the literature as a paradigmatic example of how self-locating uncertainty 'creates havoc' for standard Bayesian principles of Conditionalisation and Reflection. Furthermore, it is also thought to raise serious issues for diachronic Dutch Book arguments (see Titelbaum, 2016b, 2013). I show that, contrary to the consensus view, it is possible to represent the Sleeping Beauty problem within a standard Bayesian framework. Once the problem is correctly represented, the solution satisfies all the standard Bayesian principles, including Conditionalisation and Reflection, and is immune from Dutch Book arguments. Moreover, the solution does not make any appeal to the Restricted Principle of Indifference that is generally accepted in the literature on self-locating uncertainty, which, I argue, is incompatible with the principles of Bayesian reasoning.

7.1 THE PROBLEM

Adam Elga (2000) introduced to the philosophical literature what has come to be known as the *Sleeping Beauty problem*:

> Some researchers are going to put you to sleep. During the two days that your sleep will last, they will briefly wake you up either once or twice, depending on the toss of a fair coin (Heads: once; Tails: twice). After each waking, they will put you to back to sleep with a drug that makes you forget that waking. When you are first awakened, to what degree ought you believe that the outcome of the coin toss is Heads? (Elga, 2000, p. 143).

Elga's description of the problem grants the following natural assumptions:

1. The experiment lasts two full days, from the moment the experimental subject (Beauty) is put to sleep at the end of day 0, to the moment when she is woken up and dismissed at the beginning of day 3.

2. There are two possible outcomes to the experiment: either the coin toss comes up Heads, and Beauty is woken up on day 1, but left to sleep on day 2; or the coin toss comes up Tails, and Beauty is woken up both on day 1 and on day 2. Each outcome has a prior probability that is equal to $\frac{1}{2}$.

3. When she wakes up during the experiment, Beauty does not know which day it is.

From Beauty's standpoint, the task is to determine the probability of Heads, after she wakes up during the experiment. I will assume that Beauty is a Bayesian.

On each day, Beauty could be in either of two states: she is either awake or she is asleep. Representing an awakening by $\mathfrak{w}$ and a sleep-through by $\mathfrak{s}$, we know from the outset that day 1 involves an awakening, while day 2 may involve either an awakening or a sleep-through, depending on the result of the coin

toss. A first characterisation of the sample space for the whole experiment is therefore:

$$\Omega = \{\mathfrak{ws}, \mathfrak{ww}\}$$

Let $H = \{\mathfrak{ws}\}$ be the event that the coin toss comes up Heads and $T = \{\mathfrak{ww}\}$ be the event that the coin toss comes up Tails. By assumption 2, the prior probability of H is the same as the prior probability of T, that is, $P(H) = P(T) = \frac{1}{2}$. In the context of the experiment undergone by Sleeping Beauty, the probabilities of the events H and T are given as priors, as they are fixed by the experimental setup.

During the course of the experiment, Beauty is allowed to make some observations that potentially provide her with side information about the outcome of the experiment. Each observation consists in waking up on a given day, and noting that 'Beauty wakes up on day i', where $i \in \{1,2\}$ stands for the current day. (Recall that, by assumption 3, Beauty does not know which day it is when she wakes up, and that by assumption 1 each outcome spans over two days.) How does this observation affect the probability of H? In order to answer this question, we need to represent the observation that 'Beauty wakes up on day i' as an event within the same sample space as the event H. Let W be the event that Beauty wakes up on day i. A quick glance at Ω reveals that without some further elaborations, that space is not sufficiently rich to express the event W. This is because, if the outcome of the experiment is $\mathfrak{ws}$ (if, that is, the result of the coin toss is Heads), it is indeterminate whether Beauty wakes up on day i. To see this, consider how the outcome $\mathfrak{ws}$ consists of an awakening followed by a sleep-through, so if $i = 1$, Beauty wakes up, but if $i = 2$, Beauty does not wake up.

The difficulty with modelling the event W that Beauty wakes up on day i is due to the fact – expressed by assumption 3 – that Beauty does not know which day it is when she wakes up. To represent her uncertainty, we need to refine the outcome space, taking into account that the experiment spans over two days (assumption 1), and that, before any specific observation is made, it could be either day 1 or day 2, since the experiment lasts two days regardless of the result of the coin toss. The resulting refined sample space is:

$$\Omega' = \{\mathfrak{ws}1, \mathfrak{ws}2, \mathfrak{ww}1, \mathfrak{ww}2\}$$

In the sample space Ω', the event corresponding to a Heads result of the coin toss is $H = \{\mathfrak{ws}1, \mathfrak{ws}2\}$, while a Tails result corresponds to $T = \{\mathfrak{ww}1, \mathfrak{ww}2\}$. Moreover, $\mathfrak{ws}1$ corresponds to the outcome where the result of the coin toss is Heads, and it is day 1; $\mathfrak{ws}2$ corresponds to Heads and day 2; $\mathfrak{ww}1$ corresponds to Tails and day 1; and, finally, $\mathfrak{ww}2$ corresponds to Tails and day 2. As can be easily checked, the refined sample space Ω' allows us to consider a richer class of events than Ω. We can express what day is today, which enables us to represent Beauty's uncertainty regarding this bit of information. Let $D_1 = \{\mathfrak{ws}1, \mathfrak{ww}1\}$ be the event that it is day 1, and $D_2 = \{\mathfrak{ws}2, \mathfrak{ww}2\}$ be the event that it is day 2. Finally, we can express the event that Beauty wakes up on day $i \in \{1,2\}$ as $W = \{\mathfrak{ws}1, \mathfrak{ww}1, \mathfrak{ww}2\}$.

The same prior constraints set by the experimental setup should apply to the refined sample space Ω', as they did to the more coarse-grained version of the sample space Ω. In particular, by assumption 2, the probabilities assigned to H and T relative to Ω' should be equal. Table 2 summarises the refined sample space and the probabilities associated to each outcome, subject to the constraint that $P(H) = P(T) = \frac{1}{2}$.

Outcomes	(Heads)		(Tails)	
	$\mathfrak{ws}1$	$\mathfrak{ws}2$	$\mathfrak{ww}1$	$\mathfrak{ww}2$
Probabilities	$\frac{1}{2}\alpha$	$\frac{1}{2}(1-\alpha)$	$\frac{1}{2}\beta$	$\frac{1}{2}(1-\beta)$

Table 2: The Sleeping Beauty experiment

The parameters α and β in table 2 both take values in the $[0,1]$ interval and represent the conditional probability that day 1 is sampled, given that the result of the coin toss is either Heads or Tails; more precisely, $P(D_1|H) = \alpha$ and $P(D_1|T) = \beta$. These conditional probabilities are not fixed by the experimental setup, so the description of the problem leaves us free, in principle, to set them however seems best. Moreover, the statement of the Sleeping Beauty problem does not explicitly include the constraint that α and β be equal. It might be that Beauty should regard it as more likely that a certain day is sampled given that, for instance, the result of the coin toss is Heads than it would if the result was Tails. The description of the experimental setup does not give explicit information regarding how Beauty should apportion these probabilities, and therefore we should, at this stage of representing the problem, leave open how she sets the values of both α and β. A discussion of which are the correct or the more plausible values of α and β is left until later (see §7.2 below).

We are now in a position to formally state the problem (which we originally formulated as: *What probability should Beauty assign to Heads, given that she wakes up today?*) in terms of computing the posterior probability of H, given that W is observed, or $P(H|W)$. As I assumed that Beauty is a Bayesian, we immediately have:

$$P(H|W) = \frac{P(W|H)P(H)}{P(W)}$$

Since H and T partition the outcome space Ω', by the law of total probability we have that:

$$P(W) = P(W|H)P(H) + P(W|T)P(T)$$

Moreover, we know that $P(H) = P(T) = \frac{1}{2}$, $P(W|T) = 1$ (since Beauty wakes up every day if T) and $P(W|H) = \alpha$ (since the probability that Beauty wakes up, given that the coin toss comes up Heads, is equal to the probability that it is day 1 given H). The previous equation simplifies to:

$$P(W) = \frac{1+\alpha}{2} \tag{1}$$

The solution to the Sleeping Beauty problem is therefore given by the equation:

$$P(H|W) = \frac{\alpha}{1+\alpha} \tag{2}$$

Answering Sleeping Beauty's original question therefore depends solely on the value that we assign to parameter α.

7.1.1 *Further Questions*

The original Sleeping Beauty problem involves computing the value of H given that W is observed. But once this is done, there are many questions we can still ask. For example, what credence should Beauty assign to the coin toss having come up Heads, if after waking up she were informed that it is day 1? Or, similarly, what would her credence in Heads be if the experimenters told her, after waking her up, that today is the last time she wakes up during the experiment?

Another advantage of the refined sample space Ω' is it allows us to model these further questions. The first question (*What is the probability of Heads, if today is day 1?*) can be answered by computing the posterior probability of H, given D_1:

$$P(H|D_1) = \frac{\alpha}{\alpha + \beta} \tag{3}$$

Let $L = \{\mathfrak{ws}1, \mathfrak{ww}2\}$ be the event that Sleeping Beauty wakes up for the last time on day $i \in \{1,2\}$. To answer the second question (*what is the probability of H, if today is the last time you wake up?*), we just need to compute the probability of H, given the information that L is the case, which is:

$$P(H|L) = \frac{\alpha}{1 + \alpha - \beta} \tag{4}$$

Another question that will be interesting to consider might be posed to Beauty before the experiment actually begins: *Suppose that it is either day 1 or day 2. What is the probability that you wake up today?* In order to answer this question, Beauty should effectively state what is the prior probability of W, given that today is either the first or the second day (that is, given $D_1 \cup D_2$). Since $D_1 \cup D_2 = \Omega'$, the conditional probability of W given $D_1 \cup D_2$ is equal to its unconditional probability:

$$P(W|D_1 \cup D_2) = P(W) \tag{5}$$

Things are a bit different if the question specifies which particular day is sampled: *Suppose it is day 1 (day 2). What is the probability of W?* In this case,

since we know that $P(D_1) = \frac{\alpha+\beta}{2}$ and $P(D_2) = \frac{2-\alpha-\beta}{2}$, by a simple calculation we have:

$$P(W|D_1) = 1 \qquad P(W|D_2) = \frac{1-\beta}{2-\alpha-\beta} \tag{6}$$

Equation 6 states an interesting result. The probability that Beauty wakes up, given that it is day 1, is equal to 1 – just as we would expect, since the experimental setup specifies that Beauty always wakes up on day 1. However, the probability that Beauty wakes up on day 2 does not necessarily equal $\frac{1}{2}$, as this depends on what values α and β take.

7.2 SOLUTION

As we've seen in the previous section, the solution to the Sleeping Beauty problem comes down to computing the value of $P(H|W)$ in equation 2. To do this, we need to specify what is the value of α, which is not explicitly fixed by the description of the experimental setup. Is there a correct or most plausible assignment of value to α? Moreover, does the value assigned to α match the value assigned to β, or do they differ? (Although the value of β does not matter to the solution of the original Sleeping Beauty problem, it affects the solution to the further questions I described in the previous section.)

There are, I think, two possible ways to go from here. One possibility would be to say that we simply cannot assign any value to α and β, since these are left unspecified by the description of the experimental setup. This would leave us

unable to compute the posterior probability of Heads given that Beauty wakes up. Although theoretically coherent, this solution is not very attractive.

A second way to go is to proceed and assign a value to α (and to β, if we have any interest in answering the further questions that Sleeping Beauty might consider, described in §7.1.1). If we take this route, it remains to decide which values are the most appropriate. The most natural assignment – in my view – is $\alpha = \frac{1}{2}$ and $\beta = \frac{1}{2}$. The rationale for this assignment is the following. Since Beauty is uncertain about what day it is, we should conceptualise the day i that she observes as if it had been sampled through some sort of randomising mechanism. We don't need to be very specific about the nature of the sampling mechanism that we imagine. The point is simply that, in order to represent Beauty's uncertainty about what day it is, the way in which we model the problem must respect the intuition that both day 1 and day 2 are the possible objects of an observation. The parameter α (respectively, β) represents the prior probability that a day randomly sampled though this hypothetical mechanism is day 1, given that the result of the coin toss is Heads (respectively, Tails). In other words, $\alpha = P(D_1|H)$ and $\beta = P(D_1|T)$.

Since the experiment is expected to run over two days, regardless of the result of the coin toss, it makes sense to assume a uniform prior distribution over the two days in both scenarios (whether the coin toss comes up Heads or Tails). On this view, when considering what values to set for α and β, Beauty employs the Principle of Indifference: since she has no reason to believe, assuming that the result of the coin toss is Heads (respectively, Tails) and prior to any day being selected, that any one of the two experimental days is more likely to be sampled than the other, she should assign them an equal probability, and therefore set $\alpha = \frac{1}{2}$ (respectively, $\beta = \frac{1}{2}$).

I will discuss the Principle of Indifference in more detail in §7.3.1, but it is important to note here that my application of the Principle of Indifference differs from a related line of reasoning that is often applied to the Sleeping Beauty problem, which is known in the literature as the *Restricted* Principle of Indifference (RPI, for short). Although the Principle of Indifference entails RPI in the specific circumstances of the Sleeping Beauty problem, there is a profound conceptual difference between the two, which I will explain in more detail in §7.3.1 below. In a nutshell, RPI tells us that events that are – in a special sense, which I will explain later – *subjectively indistinguishable* given some evidence E, should receive equal *posterior* probability after learning that E. In contrast, the Principle of Indifference – as I have employed it – gives us a way to set the *prior* probabilities, in the absence of relevant evidence E.

7.2.1 *Answers*

To represent Beauty's uncertainty about which day it is when she wakes up within a probabilistic framework (on the assumption that Beauty reasons as a Bayesian agent) we have expanded the simple outcome space Ω, and then we employed the Principle of Indifference to generate Beauty's prior probabilities for the events that we defined relative to the refined sample space Ω'.

If we now plug in the chosen values for α and β to equations 1 and 2 from §7.2, we are finally able to compute the desired probabilities:

$$P(W) = \frac{1 + \frac{1}{2}}{2} = \frac{3}{4} \tag{7}$$

$$P(H|W) = \frac{\frac{1}{2}}{1+\frac{1}{2}} = \frac{1}{3} \tag{8}$$

And moreover, answering the further questions in §7.1.1:

$$P(H|D_1) = \frac{1}{2} \tag{9}$$

$$P(H|L) = \frac{1}{2} \tag{10}$$

$$P(W|D_1) = 1 \qquad P(W|D_2) = \frac{1}{2} \tag{11}$$

None of the above answers seems particularly surprising. If Beauty is told that it is day 1, she intuitively is in the same situation as someone who doesn't (yet) know the result of a fair coin toss, and this explains why we have a strong intuition that she should assign a probability of $\frac{1}{2}$ to Heads. Similarly, if Beauty learns that this is the last time she wakes up, but does not know if it is day 1 or day 2, she knows that it is Heads if and only if it is day 1, and Tails if and only if it is day 2. This means that $P(H|L) = P(D_1|L)$, and it is very plausible that both should equal $\frac{1}{2}$. Finally, the probabilities in equation 11 are simply in line with the description of the experimental setup.

7.2.2 *Tweaking the parameters*

Although the motivations I gave to support it in §7.2 are different, my numerical solution to the Sleeping Beauty problem agrees with the one put forward by Elga (2000), and which is known in the literature as the 'thirder' solution. However, Elga's original paper did not settle the answer to the Sleeping Beauty problem, as attested by a growing literature around it to this day. In this section, I review two prominent alternative solutions to the Sleeping Beauty problem. I show how both can be derived within the formal model of the Sleeping Beauty case that I gave in §1, and then critically examine the rationales that can be given for both.

Halfing

An early reply to Elga by David Lewis (2001) advocated a different solution to the original problem, which has come to be known as 'halfing'. According to Lewis, Beauty's credence in Heads should not change between the time before she is put to sleep and when she wakes up on day 1, but should stay equal to $\frac{1}{2}$. In other words, for Lewis, both $P(H) = \frac{1}{2}$ *and* $P(H|W) = \frac{1}{2}$. The rationale given by Lewis to defend this 'halfer' solution is that, upon waking up, Beauty does not learn anything new. She was aware all along that she would wake up at least once during the experiment, and therefore an awakening does not give her additional clues about the outcome of the coin toss. Given this consideration, halfers argue that Beauty should not update her credence in Heads from what it already was before the experiment. Since she knows the coin to be fair, she should maintain a credence of $\frac{1}{2}$ in Heads.

Lewis, like Elga, accepts that the sample space we should use to model the Sleeping Beauty problem is analogous to the the refined sample space Ω' that

I have given in §1 (see table 2). This is because both Lewis and Elga agree that we need to model Beauty's uncertainty regarding what day it is, since this information is relevant to the probability of Heads. They also agree (as, to the best of my knowledge, everyone in the literature) that, if the coin lands Tails, Beauty is equally likely to wake up on day 1 as she is on day 2, that is $\beta = \frac{1}{2}$. However, in order to get Lewis's solution, the conditional probability of day 1 given Heads should be set equal to 1, that is, $\alpha = 1$. When we set the parameters in this way, the probability of waking up on day i is:

$$P(W) = \frac{1+1}{2} = 1 \tag{12}$$

And the resulting numerical solution to the Sleeping Beauty problem is therefore:

$$P(H|W) = \frac{1}{1+1} = \frac{1}{2} \tag{13}$$

The halfer solution is known to generate some counterintuitive answers when it comes to the further questions I formulated in §7.1.1. These can all be easily derived in the formal model I have given in §1. One problem for the Lewisian halfers is that the probability that Beauty assigns to Heads appears to increase if, upon awakening, she is informed that it is day 1, as (by equation 3):

$$P(H|D_1) = \frac{1}{1+\frac{1}{2}} = \frac{2}{3}$$

This result is clearly puzzling, since Beauty's awakening on day 1 happens independently of the result of the coin toss. Lewis himself acknowledged the

puzzling nature of this result, arguing that it constitutes an interesting case of getting evidence 'about the future' (Lewis, 2001, p. 175).

In spite of the puzzling answer it generates, Lewisian halfing remains a relatively popular solution. This is because it makes a basic appeal to an intuition that is shared by many people, regarding what is the content of Beauty's evidence upon waking up. The idea behind Lewisian halfing is the following: if Heads, Beauty can only wake up on day 1. Moreover, Beauty only observes a day if she gets to wake up on that day. Therefore, if Heads, day 1 is sampled with certainty. There is no possible scenario in which Beauty gets to observe day 2, if Heads, and so the prior probability assigned to day 2 given Heads should be 0.

Although it appears intuitive, this motivation for the halfer solution may rest on a misunderstanding of what is the observable event W in the Sleeping Beauty experiment. As I explained in §1, W contains all the outcomes of the refined sample space Ω' in which Beauty is awake on day $i \in \{1,2\}$. When she is put to sleep, Beauty considers it possible that she will not wake up on every day during the experiment. This is because she knows that it is possible that she will sleep through day 2, if the result of the coin toss is Heads. So, even though she knows that she will not be consciously aware of it if and when it happens, *day 2 given Heads* is a live possibility at the outset, to which she intuitively should assign a positive prior probability. If the halfer solution were correct, however, then Beauty would be certain, even before the beginning of the experiment, that the prior probability of W is equal to 1, since (plugging in $\alpha = 1$ in equation 1) $P(W) = \frac{1+1}{2} = 1$. Moreover, puzzlingly, she would also be certain to wake up, conditional on it being day 2, as we can see by solving equation 6: $P(W|D_2) = \frac{1-\frac{1}{2}}{2-1-\frac{1}{2}} = 1$. In other words, according to Lewisian halfing, Beauty would be antecedently certain that she wakes up on every day during the experiment – even though this clearly is contrary to the description of the exper-

imental setup, which specifies that the prior probability that she wakes up on day 2 is equal to the probability that the coin toss comes up Tails – which, the coin being fair, is in turn equal to $\frac{1}{2}$.

Another reason to believe that the halfer solution rests on a misunderstanding is that halfers assign a value of $\beta = \frac{1}{2}$, on the basis of the same Restricted Principle of Indifference advocated by Elga (but which is not necessary to derive the thirder solution, as I have shown). This means that (at least in the case where the coin toss comes up Tails) halfers allow for the possibility that we should think of the day Beauty observes as if it were randomly sampled from the set of possible days within a Tails run. But why should this same reasoning not apply to the Heads run, as well? After all, the indifference should reflect the ignorance of which day it is according to Beauty's priors, and not be taken as a way to set her posterior probabilities.

Double halfing

As we have seen, the Lewisian halfing solution to the Sleeping Beauty problem proceeds from an intuitively plausible assumption about the content of Beauty's evidence when she wakes up (that is: *Beauty doesn't learn anything new, because she knew all along that she would wake up at least once*), but leads to some implausible conclusions. In particular, what has been considered especially puzzling about Lewisian halfing is that it predicts that Beauty's credence in Heads, when she is informed that it is day 1, is equal to $\frac{2}{3}$. In order to obviate this counterintuitive result, another solution has been proposed in the literature, which is known as 'double halfing' (Cozic, 2011).

According to a double halfer, Beauty's credence in Heads upon waking up is equal to $\frac{1}{2}$ (just as for Lewis). However, upon learning that it is day 1, Beauty's credence in Heads should also stay equal to $\frac{1}{2}$, in other words learning D_1 is

not relevant to the probability of H. The rationale behind this solution is quite evident, as it appears to reconcile two powerful intuitions: on the one hand, Beauty does not learn anything new relevant to Heads upon waking up; on the other hand, learning that it is day 1 is again not relevant to Heads.

A double halfer solution can be derived in my formal model of the Sleeping Beauty problem by setting the values of α and β both equal to 1.[54] Setting $\alpha = 1$ ensures that double halfers get the same answer as Lewis to the orginal problem, $P(H|W) = 1$ (as the conditional probability of H given W is just a function of α, by equation 2). Moreover, by setting $\beta = 1$ we also get – via equation 3 – that $P(H|D_1) = \frac{1}{1+1} = \frac{1}{2}$, as desired.

There's an obvious reason why setting $\alpha = \beta = 1$ recovers the double halfer solution. The double halfer solution cashes in on the fact that Beauty is certain to wake up on day 1, regardless of the result of the coin toss. This chimes both with the intuition that waking up should be uninformative (because Beauty is certain to wake up on day 1 anyway), and that learning that it is day 1 is uninformative (because Beauty is certain she observes day 1, learning it does not give her any information about the coin toss).

Besides giving the intuitively correct answer for $P(H|D_1)$, double halfing does not involve an application of the Restricted Principle of Indifference, and thus avoids the objection I raised against halfing that it represents the current day as if it was randomly sampled if Tails, but not if Heads. However, double halfing generates some very puzzling conclusions of its own. For instance, by equation 4, the conditional probability of H given that this is Beauty's last awakening (L) is:

$$P(H|L) = \frac{1}{1+1-1} = 1$$

54 Note that my claim here is just that it is possible to derive a double halfer solution within a standard Bayesian framework. Double halfers may, in fact, object to this way of representing their solution if they disagree with the representation of the problem I have given in §7.1-7.2.

That is, if the experimenters tell Beauty that today is the last time she is awake during the experiment, Beauty becomes certain that the result of the coin toss is Heads – which is clearly a very puzzling conclusion. And, even more troublingly, when both α and β are set to 1 and, consequently, the prior probability of day 2 is 0 (as $P(D_2) = \frac{1}{2}(2-\alpha-\beta) = 0$). The conditional probability of Heads, given that it is day 2, is undefined, as the denominator in the following equation is 0:

$$P(H|D_2) = \frac{P(D_2|H)P(H)}{P(D_2)}$$

But, clearly, if Beauty learns that it is day 2 the probability that she assigns to Heads is *not* undefined. Instead, it should simply go to 0, since the fact that she is awake on day 2 indicates to Beauty that the result of the coin toss is Tails.

7.3 MATTERS OF PRINCIPLE

In this section, I examine how the proposed representation and solution tally with three principles. The first principle (RPI) aims to establish a link between the subjective indistinguishability of some outcomes and their posterior probabilities. The second and the third principles (respectively, Conditionalisation and Reflection) concern instead the relationship between a rational agent's credences at different times. The solution I have proposed rejects RPI, but upholds both Conditionalisation and Reflection, when appropriately construed.

7.3.1 *Indifference, good and bad*

As noted in the discussion in §7.2, the solution I propose to the Sleeping Beauty problem involves the application of the Principle of Indifference to obtain the prior probability that one experimental day is sampled, given that the result of the coin toss is respectively either Heads or Tails. The Principle of Indifference is known to generate 'paradoxes'[55] when it is used to fix prior probabilities on uncountable domains, but I argued in §7.2 that these concerns should not worry us here, since the outcome space that we are considering in the case of Sleeping Beauty is only finite. Moreover, the Principle of Indifference is often criticised for being an 'un-objective' or arbitrary way of fixing priors, effectively allowing us to somehow extract precise probability values out of ignorance. This criticism must presuppose either that an 'objective' or correct way to fix priors exists, and so we should conform to it instead of appealing to indifference, or that there is no such way to fix our priors, and we should just limit ourselves to not having precise priors in this case. In other words, on this second line of thinking, if we have no access to objective probabilities, then we do not have precise probabilities at all.

In both instances, I think that this criticism is misguided. If we had access to a better approximation of objective probabilities, then of course we should use it in fixing the priors for the Sleeping Beauty experiment. However, the problem is precisely that we do not have access to anything of the sort, and we therefore must use other means to give a probabilistic representation of the problem. If we shouldn't assign probabilities to the events in the refined sample space Ω' on the basis that they are not fixed by the objective description of the problem, then we would simply be unable to give a numerical answer to the Sleep-

55 Such as, for instance, Bertrand's 'paradox', although it would perhaps be more accurate to view these cases as interesting, but not paradoxical (see Gyenis and Rédei, 2015).

ing Beauty problem – at least unless we are given some other ways of deriving posterior probabilities, when potentially relevant evidence is learned, but this evidence cannot be represented within our model.[56] The solution I have argued for, in contrast, requires no departure from standard Bayesian reasoning. For this reason, it should be preferred – at least, under the assumption that we want to model Beauty as a Bayesian reasoner.

The application of Indifference to fixing the value of parameter β, described in §7.2.1, implies that the conditional probability of day 1, given that Tails is the case, is equal to $\frac{1}{2}$. This conclusion agrees with a claim that is shared virtually across the board[57] in the literature on the Sleeping Beauty problem, which is that Beauty should equally divide her credences in D_1 and D_2, after learning that T. This is not surprising: based on the description of the experimental setup, Beauty knows that if Tails, she wakes up on both days. So, learning that Tails is the case does not intuitively favour either day.[58] This widely accepted claim is enshrined in a principle, which (Elga, 2000, following) is called the (Highly) Restricted Principle of Indifference (RPI, for short). The statement of RPI requires the introduction of some additional terminology.

56 We can express this technically when the evidence that is received is non-measurable with respect to the σ-algebra in our probabilistic model of the problem (Halpern, 2004, see). For example, this happens if we stick with Ω as the relevant sample space for the Sleeping Beauty problem and take $\mathscr{F} = \{\{\mathfrak{ws}\}, \{\mathfrak{ww}\}, \Omega, \emptyset\}$ as the relevant σ-algebra. As explained in §7.1, the evidence W cannot be expressed within this representation, as it is not measurable with respect to $\mathscr{F}$. Halpern explores some technical possibilities to model this kind of scenario, but I think that a basic problem with this approach is that it makes it unclear what *evidence* is. In standard Bayesian reasoning, a piece of evidence is conceptualised as an observable event. Non-measurable evidence, however, is not technically an event, since it is not an element of the σ-algebra.

57 To the best of my knowledge, no proposal explicitly denies this claim.

58 Double halfers, however, are in a puzzling predicament with respect to this widely accepted claim. If Beauty is a double halfer, learning that Tails is irrelevant to what day it is, since she is already convinced that it is day 1. Double halfers cannot reconcile the claim that Beauty assigns an equal probability to D_1 and D_2, after learning T, with the assignment of $\alpha = \beta = 1$ that is needed to derive their solution, unless they are willing to argue that Beauty's priors somehow change (in typically non-Bayesian fashion) between waking up and learning that W, and subsequently learning that T.

First of all, we need to introduce a distinction between what we will call *uncentred* and *centred* events, relative to a set of centred possible worlds (where a possible world corresponds to an outcome in the sample space that is used to represent the problem). Intuitively, (as I have explained in Chapter 2) a centred possible world is a complete description of the world which, in addition to all the objective features of the world, also specifies the time, spatial coordinate and identity of the agent that is at the 'centre' of that world. The sample space Ω', which we have used to represent the Sleeping Beauty experiment, is a set of centred worlds in this sense, as each of its elements specifies whether Sleeping Beauty wakes up on day 1 and on day 2 (which are objective features of the world) and, moreover, also specifies whether day 1 or day 2 is sampled, that is at which time Sleeping Beauty is located (the centred component). Some of the sample points (the centred worlds) within Ω' coincide with respect to the centre, as they place Beauty on the same day (for example, $\mathfrak{ws}1$ and $\mathfrak{ww}1$ both place Beauty on day 1). Some sample points, instead, coincide with respect to the objective component. For example, $\mathfrak{ww}1$ and $\mathfrak{ww}2$ both agree that Beauty wakes up on day 1 and day 2, although they place at a different time coordinate. A set of outcomes A is an *uncentred event* if, and only if, for any two centred worlds ω, ω' that agree on the objective component (but not necessarily on the centred component) A contains ω if and only if it contains ω'. For example, $T = \{\mathfrak{ww}1, \mathfrak{ww}1\}$ is an uncentred event, since it contains all the elements of Ω' that coincide on the objective component associated with a Heads run of the Sleeping Beauty experiment. By contrast, $D_1 = \{\mathfrak{ws}1, \mathfrak{ww}1\}$ is a centred event, since it does not contain all the elements of Ω' that agree with the objective component of each of its elements.

With this terminology in place, we can now state the Restricted Principle of Indifference: Let $A = \{\omega\}$ and $B = \{\omega'\}$ be two disjoint centred events, each containing a single centred world, such that their respective elements ω and ω' coincide with respect to their objective component, and let $E \subseteq \{\omega, \omega'\}$ be

the evidence available to an agent a. The Restricted Principle of Indifference says that if A and B are *subjectively indistinguishable* for a (in a sense that I will explain in a moment), then a should assign them equal probabilities. In other words, if A and B are indistinguishable and both $P(A|E)$ and $P(B|E)$ are positive, then $P(A|E) = P(B|E)$.

To illustrate what it means for two centred worlds to be 'subjectively indistinguishable', think of Beauty's predicament after she wakes up during the sleeping experiment. As we have seen, the outcomes in Ω' correspond to possible centred worlds and, by the definition of centred events that I have given, the evidence W that Beauty learns upon waking up is a centred event. Some elements of W, however, have the same objective component: namely, $\mathfrak{ww}1$ (corresponding to Tails and the first day is sampled, or $T \cap D_1$) and $\mathfrak{ww}2$ (corresponding to Tails and the second day is sampled, or $T \cap D_2$). Intuitively, Beauty cannot subjectively distinguish between these two outcomes, because in each case she feels as if this is the first time she wakes up (either because this is indeed the case, or because she forgot the previous awakening). As the two events $\{\mathfrak{ww}1\}$ and $\{\mathfrak{ww}2\}$ are both compatible with her evidence W, they have the same objective component, and are subjectively indistinguishable for Beauty, RPI tells us that $P(\{\mathfrak{ww}1\}|W) = P(\{\mathfrak{ww}2\}|W)$.[59]

Perhaps due to the notorious controversy that surrounds the regular Principle of Indifference, on the one hand, and the intuitive appeal of the claim that day 1 and day 2 are equally probable given Tails, on the other, virtually all literature on the Sleeping Beauty problem (and related problems) accepts the validity of RPI, which is often explicitly described as a fairly uncontroversial assump-

59 This is how Elga (2000) (p.145) puts it: 'If (upon first awakening) you were to learn that the toss outcome is Tails, that would amount to your learning that you are in either [day 1] or [day 2]. Since being in [day 1] is subjectively just like being in [day 2], and since exactly the same propositions are true whether you are in [day 1] or [day 2], even a highly restricted principle of indifference yields that you ought then to have equal credence in each.'

tion.[60] I think that this state of affairs should be questioned. As I see it, the problem with RPI is not that it generates wrong or implausible claims – which, in most cases (including Sleeping Beauty) doesn't seem to be true. But the logic it employs is flawed, particularly when RPI is meant to be a tool for Bayesian reasoning.

A fundamental idea for Bayesian reasoning is that we learn from experience. What this means is that the probabilities that we assign to events that we consider possible should reflect the relevant evidence that we have been able to collect up to now. The way in which evidence is incorporated into our overall probability assignment is by conditionalising – more on this in the next section. But this is only possible if we start with a probabilistic model, which assigns some prior probabilities to all the events and evidence that we expect might come our way. Notoriously, priors are not set in stone and revealed to us by some omniscient demon, so it is on this level that a Bayesian is to some extent open to guess work. Typically, as in the case of the Sleeping Beauty experiment, there are some constraints to how it is reasonable to set the priors: in the case of Sleeping Beauty, these constraints are given by the features of the experimental setup, which specifies that the coin is fair, and the precise sequence of awakenings and sleepy days corresponding to each outcome of the coin toss. But these constraints do not completely determine the prior probabilities of all the events of interest. In order to apply Bayesian reasoning, when the prior probability of some relevant evidence that we might get is not given as an external constraint, we then need to agree on fixing a prior. One possibility for doing this, as I have argued, is applying the Principle of Indifference. This is a legitimate application of Indifference: it just lets us formulate a prior, when other considerations do not help to settle the matter.

60 For a standard defense of RPI, see Elga (2004).

RPI, on the other hand, is not a principle to set prior probabilities, when other considerations fail. Instead, it places a constraint on the posterior probabilities associated to some particular pairs of outcomes – namely, outcomes that are both subjectively indistinguishable and coincide with respect to their objective components. This essentially reverses the logic of Bayesian reasoning: if we accept RPI, we accept that at least in some cases we do not just learn from experience; the conclusions we arrive to (the posterior probabilities) are not left open, but constrained by an external principle. To see this, consider two events A and B satisfying the conditions for RPI, and a piece of evidence E, consistent with both A and B. Suppose that you learn E: irrespective of any consideration about your priors, RPI now tells you that the posterior probability of $(A|E)$ is equal to that of $(B|E)$. In other words, the posterior probabilities of A and of B, given E, are constrained by RPI, and not left to be determined by your priors and the evidence E that you learn. While the numerical probabilities that we get by applying RPI might agree with what we would get by applying the Principle of Indifference, the logic behind the former is not sound and contradicts a fundamental aspect of Bayesian reasoning.

7.3.2 *Conditionalisation*

The Sleeping Beauty problem is generally taken to present a challenge to the principle of Conditionalisation.[61] Conditionalisation is the way in which Bayesian reasoners are expected to update their credences over time, upon learning new pieces of information. It works like this: suppose that at time t_1, you learn a new piece of evidence E (and nothing else). For any event A, the probability that you assign to A at t_1 after learning that E should be equal to the condi-

61 See (Titelbaum, 2016b, p. 667): 'The current consensus in the self-locating credence literature is that obtaining a general updating scheme for degrees of belief in both centered and uncentered propositions requires us to alter (or at least supplement) conditionalization in some way.'

tional probability you used to assign to $(A|E)$ at the time t_0 just before learning E. More formally, denoting by P_{t_0} and P_{t_1} your credences at t_0 and t_1, respectively, Conditionalisation places the following constraint on how your credences should change between t_0 and t_1, when the only thing that you learn in the interval between these two times is E:

Definition 13 (Conditionalisation)**.** $P_{t_1}(A) = P_{t_1}(A|E) = P_{t_0}(A|E)$

The question now is: Does Beauty update her credences via Conditionalisation upon waking up on day 1? It is often argued that if Beauty is a thirder, then the way her credence in Heads is updated when she wakes up on day 1 is not compatible with Conditionalisation. Elga himself makes this point:

> Before being put to sleep, your credence in H was 1/2. [...] when you are awakened on [day 1], that credence ought to change to 1/3. This belief change is unusual. It is not the result of your receiving new information – you were already certain that you would be awakened on [day 1]. (Elga, 2000, p. 146)

The upshot, for Elga, is that Conditionalisation does not always apply. In cases where an agent receives only centred evidence, his or her credences may change in ways that conflict with Conditionalisation.

In light of the analysis I have offered in §7.1, we can see how Elga's argument here cannot be right. To say that the change in Beauty's credence in Heads 'is not a result of [her] receiving new information' implies that Beauty is certain that she will receive evidence W, or – more precisely – it implies that the prior probability $P(W)$ equals 1. But, as we have seen, relative to the assignment of values to α and β consistent with the thirder solution, this is not true, because for $\alpha = \frac{1}{2}$ we have that $P(W) = \frac{3}{4} \neq 1$. In other words, if she is a thirder, Beauty

is *not* certain that she will learn W. Moreover, as I argued in §7.2, learning W *is* relevant to the probability of H.

The last sentence from Elga's quote indicates where the problem lies. When Elga says that Beauty does not receive new information, that is because she is certain of waking up on day 1. This explains why, intuitively, on day 0 she is certain that she will receive evidence W at least once in the future – namely, on day 1. Conditional on her being awake and it being day 1, Beauty's credence in Heads should indeed remain unchanged (as I also argued in §7.2), since $P(H)$ is independent of $P(W \cap D_1)$ – that is, $P(H|W \cap D_1) = \frac{1}{2} = P(H)$. However, upon waking up, Beauty does *not* learn that $W \cap D_1$. Instead, her evidence is just W, and since $P(H)$ is not independent of $P(W)$, this is relevant information upon which she should update her credences via conditionalisation. My solution allows this, and thus vindicates Conditionalisation.

Lewis's halfer solution – contrary to Elga's – does not entail a violation of Conditionalisation. Lewis simply starts from the assumption that the evidence W is irrelevant to H, and as we have seen this can be achieved within the representation I have given by setting $\alpha = 1$ and $\beta = \frac{1}{2}$. Given this setting, the prior probability $P(W) = 1$, and so Beauty is indeed certain that she will receive evidence W, which then gives us $P(H|W) = \frac{1}{2} = P(H)$, without any violations of Conditionalisation. Similarly, the double halfer solution with $\alpha = \beta = 1$ is consistent with Conditionalisation – at least, unless Beauty is informed that it is day 2, in which case (as we have seen in §7.2.2) her posterior credence in Heads is simply undefined.

Based on this discussion, we can now see that the key difference between the halfer and thirder solutions is the characterisation of the event W. For halfers, W is certain, and so learning W does not affect the probability of Heads. For thirders, on the contrary, W is not certain, and therefore learning it affects the

probability of Heads, via conditionalisation. Given these results, we can see that once the problem is correctly represented, the solution to the Sleeping Beauty problem does not challenge the validity of Conditionalisation as a principle for updating one's credences in the face of newly acquired evidence.

7.3.3 *Reflection*

Another rationality principle that appears to be violated in the Sleeping Beauty case is van Fraassen's Reflection principle (Van Fraassen, 1984). Suppose that you are a rational Bayesian agent, that you always plan to update your credences via Conditionalisation, and you do not expect to suffer any cognitive mishap that would lose you some of your previous evidence. Then let, as before, P_{t_i} denote your credences at a time t_i, and P_{t_j} denote your credences at some later time $t_{j>i}$. If you know, at t_i, that your later credence $P_{t_j}(A)$ in some event A will be equal to some real number $0 \leq p \leq 1$, then, intuitively, your credence at $P_{t_0}(A)$ should match that same value. That is, stated somewhat more formally (see Schervish et al. (2004)):

Definition 18 (Reflection)**.** $P_{t_i}(A|P_{t_j}(A) = p) = p$. (Assuming Conditionalisation and no evidence loss).

Clearly, you do not typically know what probability you will assign to an uncertain event in the future. This is because you do not generally know in advance which possible pieces of evidence you will learn in the future, and so you do not know what posterior probability you will assign to A by the time t_j. However, if you were certain that you will receive a particular piece of evidence E (and nothing more) by t_j, which would lead you to update your credence in A (via Conditionalisation) to $P_{t_j}(A) = P_{t_j}(A|E) = P_{t_i}(A|E) = p$, then it seems rea-

sonable to suppose that you should *already* have the same credence $P_{t_i}(A) = p$ at the earlier time t_i. This is indeed confirmed by the probability calculus: to be certain, at t_i, that you will receive evidence E just means that $P_{t_i}(E) = 1$, and so naturally $P_{t_i}(A) = P_{t_i}(A|E) = p$.

Despite this natural reading, the principle of Reflection has come under considerable critical scrutiny (Mahtani, 2016). The Sleeping Beauty problem, in particular, provides one instance when the principle of Reflection appears to be violated. If Beauty is a thirder, and assigns a probability of $\frac{1}{3}$ to Heads upon waking up on day 1, it seems that her prior credences on day 0, before the experiment begins, violate Reflection. This is because she knows, at t_0 = day 0, that she will receive the evidence W at t_1 = day 1. By Reflection, then, it seems that her earlier credence in H at t_0 should be $P_{t_0}(H|P_{t_1}(H) = \frac{1}{3}) = \frac{1}{3}$. Beauty's credence in Heads on day 0, however, is not equal to $\frac{1}{3}$ but to $\frac{1}{2}$, in accordance with what she knows about the experimental setup, which explicitly sets the prior $P(H) = \frac{1}{2}$. So, it seems that either the initial probability of Heads is not $\frac{1}{2}$, or Beauty's credences do not satisfy Reflection. Both alternatives seem very bad: the former flatly contradicts the setup of the problem, while the latter is inconsistent with the probability calculus, under the assumption that Beauty is a rational agent who updates her credences via conditionalisation. What can possibly have gone wrong?

The puzzle, I think, derives from the rather informal statement of Reflection, which has led us to a subtle mis-interpretation (see Mahtani, 2016; Briggs, 2009; Schervish et al., 2004). The 'event' upon which we are conditionalising in definition 18 is not, strictly speaking, an event with respect to the sample space Ω' that we have used to represent the Sleeping Beauty problem. In other words, to say that '$P(H) = r$' is an event which Beauty can conditionalise upon is technically incorrect, because '$P(H) = r$' is not a subset of Ω'. However, there might be some other technically legitimate way to recover the initial intuition that

Beauty is in some sense 'certain' that she will assign probability $\frac{1}{3}$ to Heads in the future, because she believes that she will receive evidence W at a particular time in the future. To find the solution to the puzzle, we need only look more closely into the conditions under which Beauty expects to learn W.

Given what she knows about the experimental setup, Beauty expects to receive evidence W on day 1, since she is certain that the experimenters wake her up on day 1 irrespective of the coin toss. This consideration is reflected in the prior probability $P(W|D_1) = 1$, as can be easily verified (see §7.2). So, when we say that Beauty is certain to learn W (and, as a consequence, to update the probability of Heads to $P(H|W) = \frac{1}{3}$), what we really mean is that Beauty is certain to experience an awakening *on day 1*. If at t_1 she was in a position to conditionalise on learning $W \cap D_1$, then indeed Reflection would be satisfied, as expected: $P_{t_0}(H|W \cap D_1) = P_{t_1}(H|W \cap D_1) = \frac{1}{2}$. However, Beauty does not learn $W \cap D_1$ at t_1, but only W. This explains why her credence at t_0 does not reflect her credence at t_1: that is not because she is irrational, or violates Conditionalisation in the way she updates her credences between these two times, but because at t_0 she can only be certain that she learns W *given* D_1, but the latter event is not part of her evidence at t_1. It would be incorrect to say that $P_{t_0}(W) = 1$, since $P_{t_0}(W) = \frac{3}{4}$. Therefore, Beauty is not *certain* of W at the earlier time, and she can't reflect on it.

7.4 BETS AND ODDS

It has been argued that in the Sleeping Beauty problem, fair betting odds and credences can 'come apart' (Bradley and Leitgeb, 2006; Briggs, 2010). If true, this would be a problem for Bayesians, who are committed to *probabilism* – namely, the thesis that rational credences always conform to the probability

calculus. On the one hand, if rational credences do not correspond to fair odds in betting scenarios, this undercuts a standard argument for probabilism (de Finetti, 1937). On the other hand, if fair betting odds are not based on rational credences, what are they and how can they be reliably computed? Appealing to additional evidence or side information that could be used to fix the betting odds will not solve the problem for the Bayesian, since any such evidence, if available, should also be reflected by the credence function.

As I showed in the previous sections, both the thirder and the halfer solutions can be represented within a Bayesian framework, in a way that is compatible with the principle of Conditionalisation. A well known result by Lewis (2010) (and generalised by Skyrms, 2009) shows that a Bayesian agent who plans to update her credences via conditionalisation cannot fall victim to what is called a *diachronic Dutch Book*, that is she does not expect now to accept some bets in the future that, by her own lights, guarantee a sure loss. For illustration, imagine that Betty is a Bayesian who plans to update her credences via conditionalisation. Before a fair die is rolled, we can assume that she might accept a bet X that pays £5 if the die shows a 3, and loses £1 otherwise. The expected value of this bet for Betty now is 0, as she expects to lose £1 with a probability of $\frac{5}{6}$, and win £5 with a probability of $\frac{1}{6}$. She also currently estimates that the probability that she wins the bet, conditional on the die showing an odd number, is equal to $\frac{1}{3}$, which is higher than her current unconditional probability of winning. So, Betty would also be prepared now to accept a conditional bet Y that pays £9 if the die shows a 3, and loses £3 otherwise, all conditional on the die showing an odd number (that is, the bet is void if the die shows an even number, but gives 1:2 odds on 3, conditional on an odd number). Suppose that later, the die is rolled and Betty receives the information that it shows an odd number. At that point, given that she plans to update her credences via conditionalisation, she will be prepared to accept an unconditional bet Z on the die showing a 3, at the same odds as the conditional bet Y. The expected

value of Z, later, is the same for Betty as the expected value of Y now – so, if the expected value of Y is non-negative, the same must be true for Z.

Regardless of whether she is a halfer or a thirder, on the solution I have given in §7.2 Beauty does not violate conditionalisation. So, by Lewis's result, she will not be dutch bookable, just as Betty would not be in the example I just gave. This is confirmed by checking the expected value she would assign to different bets on H, before and after waking up (that is, before and after W is observed). For instance, let us assume that Beauty is a halfer. Before the experiment begins, the unconditional probability that she assigns to W, as we have seen, is 1. So, the bets that she is prepared to accept now, are the same as those that she would be prepared to accept later, after W is observed. In particular, if she is willing now to accept bets at even odds for H *vs* T, she will also be prepared to accept the same bets later. On the other hand, if we assume that Beauty is a thirder, then observing W is relevant to H, since the unconditional probability that she assigns to W is $\frac{3}{4}$. Before the experiment begins, Beauty is prepared to accept unconditional bets at even odds for H, but conditional on W she would only accept bets that give 1:2 odds for H (see the answer to question 2 in §7.2.1). These are also the same odds that she would accept for an unconditional bet on H, after she wakes up during the experiment.

Contrary to what I just said, Bradley and Leitgeb (2006) argue that if Beauty is a halfer, the expected value of an unconditional bet at even odds for H, after she wakes up, is not 0 but a negative value. The reasoning behind this is that if H occurs, then Beauty is in a position to accept an unconditional bet on H (having observed W) only once, whereas if T occurs she is in a position to accept the same unconditional bet on H twice. So the idea is that, over the course of the experiment, she may accept a losing bet on H twice, and a winning bet on H only once. Consequently, the odds at which Beauty is prepared to accept a

bet on H should be 1:2, to reflect what is intuitively the statistical likelihood of being offered a winning bet.

Note that the motivation for betting at thirder odds that Bradley and Leitgeb give is based on statistical considerations that are entirely consistent with the thirder solution that I have advocated in §7.2.1. However, to square these considerations with the halfer solution that they favour, Bradley and Leitgeb must implicitly assume that the expected value of a bet is not equal to the probability-weighted sum of its value in all possible outcomes (which is the standard definition, implicit in all I have said so far. See Bovens and Rabinowicz, 2011; Briggs, 2010; Draper and Pust, 2008). As a result, their argument cannot be taken as a counterexample to the results of Lewis (2010) and Skyrms (2009), which are based on the standard definition of the expected value of a bet. In addition, their argument does not undermine the standard arguments for probabilism, since the reason they cite in favour of betting at thirder odds naturally supports the thirder solution that I have advocated, which is compatible with all the standard Bayesian principles.

7.5 CONCLUSION

The Sleeping Beauty problem has generated a great deal of controversy, as all the main attempts to solve it in the literature appear to violate some or other rationality constraint (Titelbaum, 2016b). The thirder solution originally put forward by Elga (2000) violates the principles of Conditionalisation and Reflection, while halfer solutions seemed vulnerable to a diachronic Dutch Book. A Restricted Principle of Indifference, generally accepted in the literature and designed to guide the updating of centred credences, goes against the spirit of Bayesian reasoning.

I have shown that it is possible to model a range of possible solutions to the Sleeping Beauty problem in a Bayesian framework. The thirder solution that I argued for is not dependent on the dubious RPI, and has the advantage of respecting both Conditionalisation and Reflection, in addition to being diachronically coherent.

I take the main lessons that can be learned from my discussion to be the following:

1. Bayesian reasoning can be naturally applied to self-locating uncertainty. We don't have to reform probability theory, or design new updating schemes to deal with this sort of cases.

2. In order to avoid puzzling conclusions, it is important to model evidence correctly. In particular we should be careful to model what is the prior probability of receiving different pieces of evidence.

3. Self-locating uncertainty does not cause a rift between fair betting odds and credences, and the case of Sleeping Beauty does not undermine Dutch Book arguments for probabilism and conditionalisation.

BIBLIOGRAPHY

Bovens, L. and Rabinowicz, W. (2011). Bets on hats: On Dutch Books against groups, degrees of belief as betting rates, and group-reflection. *Episteme*, 8(3):281–300.

Bradley, D. and Leitgeb, H. (2006). When betting odds and credences come apart: More worries for Dutch Book arguments. *Analysis*, 66(2):119–127.

Briggs, R. (2009). Distorted reflection. *The Philosophical Review*, 118(1):59–85.

Briggs, R. (2010). Putting a value on beauty. In Gendler, T. S. and Hawthorne, J., editors, *Oxford Studies in Epistemology*, volume 3, pages 3–34. Oxford University Press.

Brogaard, B. (2010). Perspectival truth and color primitivism. In Wright, C. D. and Pedersen, N. J. L. L., editors, *New Waves in Truth*, pages 1–34. Palgrave Macmillan.

Brogaard, B. (2012). Moral relativism and moral expressivism. *The Southern Journal of Philosophy*, 50(4):538–556.

Cappelen, H. and Dever, J. (2013). *The Inessential Indexical: On the Philosophical Insignificance of Perspective and the First Person*. Oxford University Press.

Carnap, R. (1950). *Logical Foundations of Probability*. University of Chicago Press.

Chalmers, D. (2006). The foundations of two-dimensional semantics. In Garcia-Carpintero, M. and Macia, J., editors, *Two-Dimensional Semantics*, pages 55–140. Oxford University Press.

Cozic, M. (2011). Imaging and sleeping beauty: A case for double-halfers. *International Journal of Approximate Reasoning*, 52(2):137–143.

de Finetti, B. (1937). Foresight: Its logical laws, its subjective sources. In Kyburg, H. E. and Smokler, H. E. K., editors, *Studies in Subjective Probability*. Robert E. Kreiger Publishing Co., Huntington, NY.

Draper, K. and Pust, J. (2008). Diachronic Dutch Books and Sleeping Beauty. *Synthese*, 164(2):281–287.

Easwaran, K. (2013). Expected accuracy supports conditionalization – and conglomerability and reflection. *Philosophy of Science*, 8(1):119–142.

Egan, A. (2006a). Appearance properties? *Noûs*, 40(3):495–521.

Egan, A. (2006b). Secondary qualities and self-location. *Philosophy and Phenomenological Research*, 72(1):97–119.

Egan, A. (2010). Disputing about taste. In Feldman, R. and Warfiled, T. A., editors, *Disagreement*, pages 247–292. Oxford University Press.

Egan, A. (2012). Relativist dispositional theories of value. *The Southern Journal of Philosophy*, 50(4):557–582.

Egan, A., Weatherson, B., and Hawthorne, J. (2005). Epistemic modals in context. In Preyer, G. and Peter, G., editors, *Contextualism in Philosophy*, pages 131–168. Oxford University Press.

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

Elga, A. (2004). Defeating Dr. Evil with self-locating belief. *Philosophy and Phenomenological Research*, 69(2):383—396.

Frege, G. (1892). Über Sinn und Bedeutung. In *Zeitschrift für Philosophie und philosophische Kritik*, volume 100, pages 25–50.

Glynn, L. (2010). Deterministic chance. *British Journal for the Philosophy of Science*, 61(1):51–80.

Greaves, H. and Wallace, D. (2006). Justifying conditionalization: Conditionalization maximises expected epistemic utility. *Mind*, 115:607–632.

Gyenis, Z. and Rédei, M. (2015). Defusing Bertrand's paradox. *The British Journal for the Philosophy of Science*, 55(2):349–373.

Hájek, A. (2003). What conditional probability could not be. *Synthese*, 137(3):273–323.

Hájek, A. (2008). Dutch book arguments. In Anand, P., Pattanaik, P., and Puppe, C., editors, *The Oxford Handbook of Rational and Social Choice*, pages 173–195. Oxford University Press, Oxford.

Hájek, A. (2012). Interpretations of probability. In Zalta, E. N., editor, *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/archives/win2012/entries/probability-interpret/.

Halpern, J. (2004). Sleeping Beauty reconsidered: Conditioning and reflection in asynchronous systems. In Gendler, T. S. and Hawthorne, J., editors, *Proceedings of the Twentieth Conference on Uncertainty in AI*, pages 111–142. Oxford University Press.

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

Joyce, J. M. (1998). A nonpragmatic vindication of probabilism. *Philosophy of Science*, 65:409–23.

Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitrechnung*. Ergebnisse Der Mathematik. Translated as *Foundations of Probability*, New York: Chelsea Publishing Company, 1950.

Lewis, D. K. (1979). Attitudes *de dicto* and *de se*. *The Philosophical Review*, 88(4):513–543.

Lewis, D. K. (1980). A subjectivist's guide to objective chance. In Jeffrey, R. C., editor, *Studies in Inductive Logic and Probability*, volume 2, pages 263–294. University of California Press, Berkeley.

Lewis, D. K. (1983). Individuation by acquaintance and by stipulation. *The Philosophical Review*, 92(1):3–32.

Lewis, D. K. (1986). *On the Plurality of Worlds*. Blackwell, Malden, MA.

Lewis, D. K. (1999). *Papers in Metaphysics and Epistemology*. Cambridge University Press, Cambridge.

Lewis, D. K. (2001). Sleeping Beauty: reply to Elga. *Analysis*, 61(3):171–176.

Lewis, D. K. (2010). Why conditionalize? In Eagle, A., editor, *Philosophy of Probability: Contemporary Readings*, pages 403–407. Routledge.

Liao, S. (2012). What are centred worlds? *The Philosophical Quarterly*, 62(247):294–316.

List, C. and Pettit, P. (2011). *Group Agency: The Possibility, Design and Status of Corporate Agents*. Oxford University Press.

List, C. and Pivato, M. (2015). Emergent chance. *The Philosophical Review*, 124(1):119–152.

Magidor, O. (2015). The myth of the *De Se*. *Philosophical Perspectives*, 29(1):249–283.

Mahtani, A. (2016). Deference, respect and intesionality. *Philosophical Studies*, 174:163–183.

Moss, S. (2012). Updating as communication. *Philosophy and Phenomenological Research*, 85(2):225–248.

Ninan, D. (2012). Counterfactual attitudes and multi-centered worlds. *Semantics and Pragmatics*, 5(5):1–57.

Ninan, D. (2016). What is the problem of *de se* attitudes? In Garcia-Carpintero, M. and Torre, S., editors, *About Oneself: De Se Attitudes and Communication*. Oxford University Press.

Pagin, P. (2016). *De Se* communication: Centered or uncentered? In Garcia-Carpintero, M. and Torre, S., editors, *About Oneself: De Se Attitudes and Communication*, pages 273–305. Oxford University Press.

Paul, L. A. (2017). *De se* preferences and empathy for future selves. In Hawthorne, J. and Turner, J., editors, *Philosophical Perspectives (Metaphysics)*.

Perry, J. (1979). The problem of the essential indexical. *Noûs*, 13(1):3–21.

Popper, K. (1959a). *The logic of scientific discovery*. Basic Books.

Popper, K. R. (1959b). The propensity interpretation of probability. *British Journal of the Philosophy of Science*, 10:25–42.

Quine, W. V. (1968). Propositional objects. *Critica: Rivista Hispanoamericana de Filosofia*, 2(5):3–29.

Ramsey, F. P. (1931). Truth and probability. In Braithwaite, R. B., editor, *The Foundations of Mathematics and other Logic Essays*, pages 156–198. Harcourt, Brace and Company, New York.

Reichenbach, H. (1949). *The Theory of Probability*. University of California Press, Berkeley.

Savage, L. J. (1954). *The Foundations of Statistics*. Wiley, New York.

Schervish, M. J., Seidenfeld, T., and Kadane, J. B. (2004). Stopping to reflect. *Journal of Philosophy*, 101(6):315–322.

Shaffer, J. (2007). Deterministic chance? *British Journal for the Philosophy of Science*, 58(2):113–140.

Skyrms, B. (2009). Diachronic coherence and radical probabilism. In Galavotti, M. C., editor, *Bruno de Finetti, Radical Probabilist*, volume Texts in Philosophy, 8, pages 85–96. College Publications, King's College London.

Staffel, J. (2015). Measuring the overall incoherence of credence functions. *Synthese*, 192(5):1467–1493.

Stalnaker, R. C. (1981). Indexical belief. *Synthese*, 49:129–151.

Stalnaker, R. C. (2008). *Our Knowledge of the Internal World*. Oxford University Press.

Stalnaker, R. C. (2011). Responses to Stoljar, Weatherson and Boghossian. *Philosophical Studies*, 155:467–79.

Stalnaker, R. C. (2014). *Context*. Oxford University Press.

Stalnaker, R. C. (2016). Modeling a perspective on the world. In Garcia-Carpintero, M. and Torre, S., editors, *About Oneself: De Se Attitudes and Communication*. Oxford University Press.

Stojanovic, I. (2016). Speaking about oneself. In Garcia-Carpintero, M. and Torre, S., editors, *About Oneself: De Se Attitudes and Communication*. Oxford University Press.

Titelbaum, M. G. (2008). The relevance of self-locating beliefs. *Philosophical Review*, 117(4):555–606.

Titelbaum, M. G. (2013). Ten reasons to care about the sleeping beauty problem. *Philosophy Compass*, 8(11):1003–1017.

Titelbaum, M. G. (2016a). *Fundamentals of Bayesian Epistemology*. Unpublished manuscript.

Titelbaum, M. G. (2016b). Self-locating credences. In Hájek, A. and Hitchcock, C., editors, *The Oxford Handbook of Probability and Philosophy*. Oxford University Press.

Van Fraassen, B. C. (1984). Belief and the will. *The Journal of Philosophy*, 81(5):235–256.

Venn, J. (1876). *The Logic of Chance*. Macmillan, London. (Second edition. Reprinted, New York: Chelsea Publishing Co., 1962).

von Mises, R. (1957). *Probability, Statistics and Truth*. Macmillan, New York. (revised English edition).

Weber, C. (2015). Indexical beliefs and communication: Against Stalnaker on self-location. *Philosophy and Phenomenological Research*, 90(3):640—-663.

Williams, J. R. G. (2012). Generalized probabilism: dutch books and accuracy domination. *Journal of Philosophical Logic*, 41(5):811–840.

Williamson, T. (2000). *Knowledge and its Limits*. Oxford University Press.