The University Of Sheffield.

The Experimental Approach to Vagueness

By:

Joshua James Matthews

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Sheffield
Faculty of Arts and Humanities
Department of Philosophy

May 2018

Abstract

This thesis is about philosophical theorising about vagueness, and the possibility of contributions from experimental work. I defend the view that empirical studies can inform theorising, given that the philosophical project is an essentially descriptive one, and I aim to provide a better theoretical framework to interpret empirical results than hitherto seen in the literature.

In the first half of the thesis (chapters 1-3), I introduce vagueness and make a case for the experimental approach to it, but argue that some recent studies are unsatisfactory in their methods. Following a general introduction to the phenomena of vagueness and some theories of vagueness (chapter 1), I argue that empirical work can inform theorising about vagueness, at least *in theory* (chapter 2). I argue that if theories of vagueness are descriptive, i.e. aim to model features of real-life language use, then there is a role for experimental work, and then argue that the debate shows that philosophers *are* engaged in a descriptive project. I then argue that even though experiments can be a useful philosophical tool in theory, recent studies haven't provided significant results, and the authors of studies have failed to establish their conclusions (chapter 3). One reason for this is that the approach lacks a strong theoretical basis for the connection between any particular theory of vagueness, and particular empirical result; there isn't a strong enough principled reason for associating one theory or another with particular patterns given by participants.

In the second half of the thesis (chapters 4-6) I aim to make progress with supplying this theoretical back drop. I first argue that logic and semantic theories, such as those that come with a theory of vagueness, restrict *theories of mind*; a theory of rational belief and decision making (chapter 4). This provides the initial justification for associating a theory of vagueness with patterns seen in experiments; it can be claimed that a particular theory gives a good explanation of an empirical result, based on that theory's commitments in the theory of mind. I then go on to discuss what the particular commitments for supervaluationism (chapter 5) and degree theoretic semantics (chapter 6) look like. I argue that in each case there is more than one consistent way to work out a theory of mind, but that the best models are a suspension of judgement model for supervaluationism, and a kind of indeterminate belief for

degree theoretic semantics. In chapter 7, I sum up the conclusions of the thesis, and consider some ways the experimental approach to vagueness might continue to make progress.

Acknowledgements

First, I'd like to thank my supervisors Rosanna Keefe and Robbie Williams. Each waded through enormous amounts of draft material, providing insights and discussion that were essential to the development of my work. Without Rosanna's own research I likely would not have started on vagueness, and her continued support and feedback over these years has been invaluable to the completion of this thesis.

I'd like to thank all those who made Sheffield a great place to study. The graduate community was supportive and friendly, and was the source of many fruitful discussions from which I benefitted. In particular, the friendship of Josh Clack, Mike McKinney, Harry Pickard and Gareth Roddy made my time in Sheffield the great experience it was.

I would like to thank my wonderful family, who helped make 8 consecutive years as a student possible. In particular each of my grandparents and my father, Matt, for emotional and financial support, and especially my mother, Katrina, to whom I am indebted for any positive quality I possess.

I am very grateful for the financial support I received from *WRoCAH*, which allowed me the privilege of embarking on graduate research.

Finally, special thanks to Grace, who has been there since the start of this journey, and has inspired me every step along it.

This work was supported by the Arts & Humanities Research Council (grant number AH/L503848/1) through the White Rose College of the Arts & Humanities.

Table of Contents

Introduction

Vagueness as a phenomenon has been well explored in the philosophical literature, and there is no shortage of candidate theories to model it or handle the paradox that comes with it. These theories standardly aim to give the semantics and logic of vague discourse, explain the features of the phenomena of vagueness and solve the sorites paradox. With several options available, how are we to decide between these theories? In this thesis, I will argue that an important part of the answer could be based on empirical data; the results of experiments designed to test the intuitions and language use of 'the folk'. Of course, the use of experimental data in philosophy is a controversial issue, and considering if some set of data should influence philosophical theorising requires a very careful approach in any area. However, I will argue that vagueness is certainly a topic in philosophy for which empirical results of the *right sort* can provide genuinely useful insights for philosophers.

I argue that this is because we should think of theories of vagueness as *models* of a real-life phenomenon of natural language, and thus as having an importantly descriptive aspect. In order to be able to successfully model this phenomenon we need to have a good idea of what it is like. This is what experimental work proposes to do; to give us a more accurate picture to model, a better understanding of what features our logic and semantic theory should have. There are rarely knock-down arguments against theories of vagueness. Rather the debate often takes the form of a 'reflective equilibrium', as it has been called. This requires weighing up what intuitions a theory retains against the intuitions it doesn't, and deciding what theory gives us the greatest benefits for its costs. Philosophers might disagree about many things during this process, including which intuitions are most important for a theory to retain, or even which are the intuitions that should be accommodated at all. The experimental approach can aim to contribute at this point, investigating the features of speaker's use of vague language and mental representations of vague concepts, in turn revealing which intuitions a theory of vagueness should attempt to model.

There are of course difficulties for this approach. A host of problems stand before a philosopher of language or logic who hopes to use evidence from an experiment to inform a theory. There are general questions that need to be addressed. A key example I will address here is the 'expertise objection': why should we be interested in the opinions of those who aren't experts in the philosophy of language or logic? There are also more specific questions

to address, that arise when reviewing contemporary contributions to the literature that do draw on empirical results. Some of these are practical issues; relating to the method used, extraneous variables affecting results and appropriate interpretations of results. Some of these are theoretical issues; questions about how to fit results into our theoretical theorising, and what type of results would support one theory of vagueness or another. I will suggest that both sorts of issue affect some recent studies, but will focus on one particular theoretical short-coming. Using experiments to test intuitions about vagueness requires recording responses to certain questions or stimuli, perhaps judgements about borderline cases, sharp boundaries or a sorites series. A simple example is recording how likely participants are to assent to a contradictory description of a borderline case, e.g. to 'red and not red' for some borderline red patch. But for such results to genuinely inform theorising, we need to know which response types support which theories, i.e. what a theory of vagueness might predict about how people do respond to borderline cases etc. I'll argue that some recent studies make somewhat simplistic assumptions about this, but it mustn't be taken for granted. A typical theory of vagueness will tell us about the logic and semantics of vague language, but isn't obvious how this translates to beliefs and attitudes, or that it does translate at all. What should we think a theory that introduces a three-valued semantics for instance, predicts about the folk's intuitions about vagueness? Why should we think it does make predictions at all, opposed to just settling the questions of logic and semantics for vague discourse whilst remaining silent about any questions of cognition or psychological states?

For the experimental project in question to get off the ground it must be clear that there *is* some connection to found between a logic and semantics, and beliefs and attitudes, that one constrains the other in some way. Following that, it needs to be established how theories of vagueness do constrain beliefs and attitudes. Whilst some philosophers do discuss connections between logico-semantic theories and cognitive theories, they have been neglected in the experimental literature. I argue that these connections should play a central role, and are important for making progress. The connections needed to get the experimental project going do exist; I think notions of 'truth' and 'logic' are necessarily cognitively loaded, and become somewhat mystical if their cognitive significance is denied. Once this has been established, we're in a position to consider what theories of vagueness do say about beliefs and attitudes.

The first part of the thesis will aim to establish that vagueness is a topic that can benefit from empirical investigation, and review some studies that have attempt to do this. In chapter 1 I

will discuss vagueness in general, and some of the theories of vagueness that will be important in this thesis. In chapter 2 I will defend the relevance of empirical data to the study of vagueness, arguing that theorising about vagueness is really trying to model a phenomenon, and that we should inform this modelling with data of what the phenomenon is like. In chapter 3 I will examine some recent studies that attempt to gather relevant data to support certain theories, arguing that none present very compelling cases, and in particular take for granted to link between formal accounts of the logic and semantics of vague language and views about psychological attitudes given that account.

The second part of the thesis will aim to establish that there is such a link, and undertake the project of examining the link for some important logico-semantic frameworks for vagueness. Chapter 4 will consider the different possible strengths of connection between logico-semantic theories and cognitive theories, and defend what I will identify as a medium-strength version of it. Chapter 5 will consider what psychological models can be associated with a supervaluationist account of the logic and semantics, suggesting that there is more than one possible version, but that a 'suspension model' is the one that is best. Chapter 6 will consider psychological models that could be associated with degree semantics. The second part then has interest beyond just its utility for experimental approaches to vagueness; it argues that any logic or semantic theory has cognitive consequences, and shows how this can play out for two different semantic approaches.

This thesis will aim to establish that empirical data can make a genuinely interesting contribution to the philosophical data on vagueness, and when studies are done right, should be paid attention. It will lay out a framework for thinking about the cognitive commitments of a theory of vagueness, and present versions of the cognitive commitments for two popular semantic approaches to vagueness.

Chapter 1 (*Another)* Introduction to Vagueness

If there were only one introduction to vagueness in the literature, then there would surely not be *too many*. And one might think it is plausible to say that if whatever number of introductions there currently are is not *too many*, then the addition of another single introduction to the literature cannot make it that there are too many. However, if we were to keep adding to this number, and more and more introductions to vagueness were to be published, to the point that there were a thousand or more, we might think we had reached the point where there *were* too many. So where did we go wrong?

This is of course a play on a familiar puzzle, the sorites paradox. It is clearly true one introduction to vagueness would not be too many, and it looks plausible that the single addition of an introduction could not take us from there not being too many in the literature to there being too many in the literature. Assuming that it is plausible that some number *would* indeed be too many, perhaps one thousand publications in the literature outlining the features of vague language and the sorites paradox, we are left with a problem, caused by the vagueness of 'too many'. This is a playful example, whether there can be too many introductions or not likely depends on the contribution each makes. But it reveals the general structure of a sorites paradox; whereby plausible assumptions about the applicability of some word or phrase lead us to a false conclusion. I hope that the reader doesn't think there are yet too many introductions to vagueness, as I am about to present another.

This chapter aims to clarify what vagueness is, and review some important approaches to modelling it. I will discuss the main features of vague language, and will then move on to details of some theories of vagueness that will be important for this thesis. Again, I won't be defending any particular theory here, but aim to outline some theories and what they attempt to do before setting out on the main aim of the thesis: the defence and development of the experimental approach to vagueness.

1. Vagueness

Let's begin with a slightly less contrived sorites paradox. Whilst sitting in the student union one day, a graduate student began to wonder if she could afford to buy a beer. She checked her bank balance, and seeing that her stipend had come in that day, she realised she was now

rich. She then reasoned that buying one beer could never turn a rich person into a person who isn't rich. It's absurd to think that spending three pounds on a beer could turn a person from rich to not rich. Thus, knowing that she was rich, she knew she'd always safely remain rich after buying a single beer. After finishing the first beer, she wondered if she could afford another. She recalled the principle which was appealed to before buying the first beer: if someone is rich, then the purchase of one beer cannot turn them not rich. As she was rich before, and had only bought one beer, she knew she must still be rich. That same principle then showed her that she could afford to buy one more. She proceeded to repeat this process all day, knowing that no single beer would be enough to stop her from being rich. Her reasoning seemed solid. However, to her horror, she checked her bank balance after a long day drinking, and realised that she had in fact become a person who isn't rich. Where did she go wrong? It was true that she was rich at the beginning of the day, and it certainly seemed true that buying any single beer could never change that fact. Yet mysteriously she was no longer rich after all.

This is an example of a sorites paradox. It takes an undeniable application of some word/phrase, and the plausible principle that there is some possible change in the relevant object small enough that it couldn't make a difference to the applicability of the word/phrase, and then through repeated applications forces us to an absurdity. Similar cases can be formed for many terms in natural language. A standard PhD thesis might be thought to be a rather long piece of work, and whilst one word could never be a big enough difference between a long piece and a not long piece, the repeated removal of single words from any long thesis will certainly eventually leave you with a piece that is not long. One slice of cake couldn't turn a skinny person into a not skinny one, yet eating a single slice enough times certainly will. A single second isn't enough time to take someone from being not old to being old, yet the repeated passing of seconds eventually delivers someone old. The reason that this paradox arises for these words is that they are *vague.* Other paradigm cases include 'red', 'young', 'child', 'small', 'heap', 'tall', for which similar puzzles can be constructed. As is often the case in the literature, I will generally use predicates as examples, but it should be noted that vagueness stretches to other types of expression, such as adverbs ('slowly') and quantifiers ('many').

These puzzles can be formalised in different ways, but I favour to do so in a way such as the following:

Let '*n'* be any natural number and 'Ra*n*' be 'person a, possessing *n* pounds, is rich'.

P1. Ra*1000000000*

P2. ∀*n*(Ra*n* → Ra*n-1*)

Therefore:

C. Ra*1*

We start with an undeniable instance where a predicate applies; premise one, a person in possession of one billion pounds is rich. We then take an intuitive conditional; premise two, which says that the subtraction of a single pound from a rich person's fortune will always leave you with a rich person. This is known as the *inductive premise*. We can then conclude that a person in possession of one pound is rich. Arguments of the same structure can be formed for any of the paradigm examples I considered above. Accordingly, we can use a sorites paradox to conclude that anyone of any age is young, and old, or that anyone is tall, and that they are not tall, or that every colour is red. We need to resist these conclusions, but the problem is that we can reach them from the apparently true premises of the sorites paradox.

The paradox can be constructed for any words or phrases that are *vague*, and it is part of the goal of philosophical theorising about vagueness to show why the paradox is unsound. After all, normal speakers communicate just fine with apparently paradoxical words like 'rich' and 'old', so there is presumably something about their meaning, that follows from the logic and semantics of vague language, that shows why the sorites paradox isn't actually sound.

1i. What makes a word 'vague'?

What does it mean for a word to be *vague*? I cannot give a straightforward response to the question of what vagueness is, as we'll soon see, the answer varies with different theories of vagueness. However, we can point out pre-theoretic features that vague words exhibit:

Possibility of borderline cases: Our vague language standardly gives rise to what are known as borderline cases; these are those cases for which it is unclear whether some word applies or doesn't apply. Whilst I am rich in possession of one million pounds, and clearly am not rich in possession of one pound, there could be amounts of money such that if I were to possess precisely that amount it would not be clear whether to count as 'rich' or not.

Similarly, someone would be tall at 7ft, and not tall at 4ft, however when we come across a person at 5'11", we might find it difficult to say whether they should be counted as tall or not tall. These are instances of borderline cases of 'tall' and 'rich'. On the spectrum between the clearly rich and clearly not rich, and between the clearly tall and clearly not tall, there are cases that appear to fail to straightforwardly fall into either category. Competent speakers wouldn't be able to easily categorise these cases one way or the other, and if forced to wouldn't widely agree on the categorisation. In these cases, it seems no further information, i.e. about how much money I have, and the distribution of wealth in the country for instance, would make the decision easier, and neither is it the case that the speakers are deficient in their understanding of the language. Our standard usage of vague language is such that borderline cases are possibilities.

It seems there are instances of borderline cases for any vague term; shades between red and orange where we wouldn't want to apply either term, lengths of essay which we wouldn't want to claim are either short or not short. Contrast this with a precise term, e.g. acute; there are no borderline 'acute' angles, an angle which one couldn't judge as either acute or not acute once we had all the relevant information and if we understood the meaning of 'acute'.

Apparent lack of sharp boundaries: Vague language also seems to lack precise extensions. 'Rich' fails to delineate a precise set of people. There is also no number of pounds such that in possession of that number one would be rich, yet would one would not be rich with a pound less. There is no way to set the exact number of pounds, such that with that amount or more, someone is 'rich' (relative to any society or economy). Similar things can be said about all of the terms above; there is no height such that one would be tall at that height, but not tall at a millimetre shorter, no number of words which make a long PhD thesis, but would make a not long thesis at that number subtracted by one. These words fail to set precise boundaries to their extension, we cannot clearly separate the tall from the not tall, the long theses from the not long ones.

Again, in this case it at least seems that no more information about the meaning of 'rich' is required, or information about the distribution of wealth in the relevant context. It's simply the case that fully understanding 'rich' isn't sufficient for being able to specify a precise range of values where 'rich' applies. In this way, vague language seems to lack sharp boundaries. Compare to this to a precise word: someone who fully understands 'acute' can

easily specify a precise range of values where ‘acute’ does apply, and a degree such that an angle of that degree is acute, but if it were one degree greater it would not be acute.

Sorites susceptibility: As discussed, vague language tends to be susceptible to *sorties paradoxes*, such as those outlined at the beginning. For this to be true of a word it just has to be the case that there is a possible positive instance of that word, i.e. a rich person, and that there is a plausible inductive premise for that word; if one is rich with *n* pounds, then one is rich with *n-1* pounds. These two plausible premises together drive you to an absurdity: that someone with only a single pound is rich.

This is true of all the vague language we’ve considered so far, sorites paradoxes could be constructed for ‘tall’, ‘long’, ‘old’ etc… Upon understanding the meaning of these vague words, it seems that the relevant inductive premise of a sorites becomes plausible. Of course, for a precise word, a sorites paradox couldn’t arise. When one understands ‘acute’, one would never find a relevant inductive premise plausible.

Tolerance: It has also been suggested that ‘tolerance’ is an essential feature of vague language (e.g. Wright 1976, Cobreros et al 2012). For a term to be tolerant is for there to be degrees of possible change small enough in a such that it wouldn’t make any difference to the applicability of a term. For example, there is a number of words (one word, for example) we could subtract from the number of words in a *long* thesis, such that wouldn’t affect the applicability of ‘long’ to that thesis. The same can be said for other words considered here; there are small enough measures of height (a millimetre) that changing by that measure couldn’t affect the applicability of ‘tall’ to someone, there short enough amounts of time (a second) such that any age changing by that amount couldn’t affect the applicability of ‘young’ to someone.

These are the typical features of vague language. I take the most essential features to be the possibility of borderline cases, and the apparent lack of sharp boundaries. These two are closely related, as noted by Keefe (2000: 7). Having borderline cases can be thought *necessary* for lacking sharp boundaries; if it isn’t possible to delineate a precise range of applicability for a word, then there will be possible borderline cases. If there seems to be no particular lower limit to ‘long’, i.e. a particular number of words such that a thesis containing that number would be long but with a single word less would not be long, then as cases drift further from the clearly long, it becomes unclear whether they are long or not. A thesis of a hundred thousand words is long, but without a clear lower limit, it is possible to have cases

that cannot be straightforwardly classified as long or not long: borderline cases. Conversely, the possibility of borderline cases is not *sufficient* for an apparent lack of sharp boundaries. If *long** were true of every thesis over 85,000 words, and false of every thesis under 80,000 words, then we have a range of borderline cases *long**: those between 80,000 and 85,000 words. However, *long** does not lack sharp boundaries, there is a precise value which is the shortest a thesis can be whilst being truly *long**, and a precise value which is the longest a thesis can be whilst being falsely *long**.

Sorites susceptibility and tolerance are also closely related. We can construct sorites paradoxes for vague language because of the plausibility of the inductive premise, which is the same as the plausibility that vague language is tolerant: the inductive premise states that a certain degree of change of some factor isn't sufficient to change the applicability of a relevant predicate, which is the same as saying that predicate is *tolerant*. Thus, sorites susceptibility is sufficient for tolerance. It is not perhaps necessary though, if we consider predicates that have no false or no true instances. If *long*** applies to any thesis containing 0 words or more, then *long*** is tolerant: a single word would never be a degree of change great enough to stop a thesis as counting as *long***. However, it is not susceptible to a sorites paradox, as its tolerance cannot drive us to an absurd conclusion. It can take us to the conclusion that a thesis containing one word is *long***, but that is true, so no paradox arises.

Here we have a picture of the primary features of vague language. It is language the competent usage of which allows for possible borderline cases, an apparent lack of sharp boundaries, and sorites susceptibility (and thus, *tolerance*). In contrast, the competent use of precise language draws sharp boundaries, and allows no borderline cases. For extra clarity, I will point out some features of language that might be called 'vague', but should be kept distinct for my purposes here:

Ambiguity: The sentence 'banana flies like fruit' might be called *vague*, as it is quite unclear. There are two possible distinct propositions that could be expressed by that sentence; one about the manner in which bananas fly, and another about the culinary tastes of banana flies. However, this is a structural ambiguity, not vagueness. The unclarity is a result of the different possible propositions expressed by the same sentence, not about the lack of sharp boundaries of any of the words in the sentence.

Context Dependence: One might call 'John is tall' vague, as its truth value can vary depending on context. It might well be true that John is tall in the context of six-year olds, but

put him in the context of professional basketball players, and it may well become false that he's tall. 'Tall' is context dependent; what is or isn't tall is relative to contexts and comparison classes. But this isn't vagueness in the sense I'm interested in. We can see that they're separate phenomena, as any unclarity about 'John is tall' based on its contextual sensitivity is removed by placing him in a context. We may not know how to judge whether John is tall until we know what context we are judging him in. Fix the context and that unclarity is removed, yet the vagueness of 'tall' remains. Even fixed in the context of professional basketball players, 'tall' fails to delineate a precise extension; we still cannot pick out the shortest height one could possibly be whilst remaining tall.

Under-specificity: If I were asked how many students studied at my university and I were to reply that there are over two hundred students, one might reasonably call my answer 'vague'. It might be true, but it leaves it quite unclear how many students there actually are at my university. This is different from the sort of vagueness I'm interested in here. What I have claimed is under-specific, it is not very informative. However, there *is* a precise number of students who study at my university, there is no unclarity due to a lack of a precise extension.

This should leave us with a clearer view of what I mean by 'vague'; a word is vague when the competent use of it allows for borderline cases and an apparent lack of sharp boundaries. I don't mean that it is (only) ambiguous, context dependent, or under-specific.

1ii. Theories of vagueness

What is the challenge posed by vagueness? The primary goal is to account for the *logic* and *semantics* of vague language, given the features just discussed. A lack of sharp boundaries and borderline cases are not straightforwardly accommodated in a classical logic and semantics. Classical logic assumes bivalence, that every proposition is true or false. But if our language lacks precise extensions, then we cannot straightforwardly categorise every proposition as true or false. Given possible borderline cases of 'rich', we may be inclined to deny that every instance of 'x is rich' is either true or false. The challenge of vagueness is to give the semantics and logic of vague language, and explain these phenomena; why does vague natural language allow for borderline cases and appear to lack sharp boundaries, compared to precise language?

The suggested logic and semantics should also give a solution to the sorites paradox. As discussed, given the general success of natural language users, we can assume that the

language is not truly inconsistent, and thus that the sorites paradox is not sound. If this is right, then one of the premises in the argument must fail to be true, or the inference must be invalid (i.e. *modus ponens* must fail). A key task of philosophers working in the area is to provide a solution to the sorites paradox, whilst offering a compelling candidate for the semantics and logic of vague language.

Many *theories* of vagueness have been proposed in the literature, which aim to do just this; provide the logic and semantics of vague discourse, explain the phenomena of vagueness and solve the sorites paradox. Here I detail a few approaches that will be relevant to the discussion in this thesis:

Supervaluationism: Supervaluationism is sometimes described as the most popular theory of vagueness, and well-known defences include Keefe (2000), Fine (1975) and Varzi (2001). This theory makes use of a notion of 'admissible precisification'; an admissible precisification is a way of making a word precise, drawing a sharp boundary in that word's extension. Vague words are those whose usage fails to delineate only one admissible precisification, but leave it unclear between a range. For instance, 'rich' fails to pick out a particular number of pounds (in a given context) such that it is least one can possess whilst being rich. Supervaluationism claims that rather 'rich' picks out several equally good candidates for this number, different places we could draw the line between the rich and the not rich, different ways of admissibly precisifying 'rich'. Precise language on the other hand picks out only one admissible precisification, there is only one admissible way of drawing a line between the acute and not acute angles.

There is more than one way of cashing out a logic and semantics at this point (as I will discuss in depth in chapter 4), but 'supervaluationism' is usually used to refer to a theory that suggests that to get the semantics for vague language we need to take all of these admissible precisifications into account: truth is truth on all admissible precisifications (supertruth), falsity is falsity on all admissible precisifications (super-falsity). Propositions that are true on some but not all admissible precisifications are *indeterminate.* Take again 'rich'. Being vague, 'rich' leaves it indeterminate between a range of equally good candidates where the boundary is between the rich and the non-rich. Each of these admissible precisifications give us a classical model, relative to any one in particular, everyone is either rich or not. The suggestion is that someone is truly rich iff they are rich according to every admissible precisification; they'd be rich regardless of where the boundary was drawn. Similarly,

someone is only not rich if it's false on all admissible precisifications that they are rich; they come out as not rich regardless of where the boundary is drawn. Cases which are rich on some admissible precisifications, but not others are indeterminately rich.

This framework captures the features of vagueness discussed above. Borderline cases of vague language are *indeterminate.* A borderline case of rich is such that it's neither true nor false that they're rich, they are rich on some but not all admissible precisifications. An apparent lack of sharp boundaries in vague language is the result of the range of admissible precisifications. There is no boundary between the rich and the not rich, because vague language doesn't pick any particular boundary out, but leaves it indeterminate between a range of boundaries. None of these boundaries could be considered *the* sharp boundary in a vague word's extension, because it is true that each is the boundary on only *one* admissible precisification (i.e. the admissible precisification that places that particular boundary), and false on the others. Sorites susceptibility and tolerance can be explained by the same point. Vague language seems sorites susceptible and tolerant as there is no *particular* boundary picked out by its usage. Thus, the inductive premise of the sorites paradox looks plausible: there is no *particular* number of pounds such that someone that number would be rich, but wouldn't be with a pound less.

Despite this lack of a counter example to the inductive premise, supervaluationism still solves the sorites paradox by denying it. On every admissible precisification, it is *false* that if someone is rich, they'd still always be rich with a pound less, as every admissible precisification picks out a sharp boundary between the rich and the not rich. As falsity is falsity on all admissible precisifications, the inductive premise comes out false. It is also true that there is a number of pounds such that someone with that number is rich, but wouldn't be with a pound less, as it is true on every admissible precisification. Thus, the sorites paradox comes out unsound, as one of its premises is false.

Whilst the inductive premise is false, there is no falsifying instance, as there is no particular number of pounds which is the least one can possess whilst being rich on *every* admissible precisification; this number is different according to each one. Instead we get a range of equally good candidates. Similarly, the claim that there is such a number is true, but lacks a verifying instance. These results show the deviation from classical semantics on supervaluationism. As has already been noted, supervaluationism denies bivalence, allowing some propositions are indeterminate, but the semantics are also not truth-functional. An

existential generalisation can be true, but lack a particular instance it is true of. It's true that there is a number of pounds which is the least one can possess and remain rich, but there is no particular number this is true off. Another interesting case of this is true disjunctions that lack a true disjunct. As on every admissible precisification, everyone is either rich or not rich, the law of excluded middle is preserved. However, as we know, some people can be indeterminately rich. Thus, for a borderline case of richness S, 'S is rich or S is not rich' is true, even though neither 'S is rich' nor 'S is not rich' is true.

Even though supervaluationism deviates from classical semantics, it is usually claimed that it preserves classical logic. There is more than one way of cashing out a consequence relation for supervaluationism, but usually it is taken to be preservation of supertruth (see Varzi 2007 for discussion of other possibilities). This preserves all classical tautologies and inferences in a standard object language. But supervaluationists often make use of a 'definitely' operator; DP is true where ever P is supertrue. But it had been shown that many classical inferences fail in cases where the definitely operator is in use (see Williamson 1994, chp. 7 and Keefe 2000, chp. 7 for discussion).

Many-valued theories: Another common type of approach is to introduce more truth-values in between truth and falsity, with which to model the fuzzy borderline cases in between the clear positive and negative cases of a vague word. Three-valued logics (e.g. Tye 1994, Hyde 2008) introduce a third value in between truth and falsity (standardly *indeterminacy*), which borderline cases take. Here the semantics remain truth-functional, unlike supervaluationism; the truth value of complex sentences is a function of the truth values of their atomic parts. So, as some propositions can be indeterminate, the law of excluded middle fails. When P is indeterminate, P v ¬P is also indeterminate. Thus, these logics are *paracomplete*, which might be thought appropriate in the context of vagueness; unlike supervaluationism, these accounts don't entail that it's true of everyone that they're 'rich or not rich', even if they're a borderline case of richness.

A three-valued theory solves the paradox by denying the inductive premise is true. On Tye's semantics (1994), which uses the matrices of the strong Kleene logic *K3,* a conditional with an indeterminate antecedent and consequent is also indeterminate. It is indeterminate that for all number of pounds, if someone is rich with some number of pounds, they'd still be rich with a pound less, as there are some instances of that conditional with indeterminate antecedents and consequents: '$\forall n$(Ra_n → Ra_{n-1})' is indeterminate, as some instances are

indeterminate, i.e. those where Ra*n* and Ra*n-1* are indeterminate. Thus, the sorites paradox comes out unsound. Hyde's semantics (2008) who follows Lukasiewicz's semantics in a three-valued model, *L3,* gives a similar result, but there a conditional with an indeterminate consequent *and* antecedent is true. However, a conditional with a true antecedent and an indeterminate consequent is indeterminate, and thus Hyde concludes that ∀*n*(Ra*n* → Ra*n-1*) is also indeterminate, as there will be an instance where Ra*n* is true and Ra*n-1* is indeterminate (2008:185-6).

The truth-functional conditional in three valued systems gives some unattractive results in a three-valued semantics. The semantics of *K3* means P → P comes out indeterminate where P is indeterminate. Where S is borderline rich, 'if S is rich then S is rich' is then indeterminate, which looks wrong. Hyde's *L3* can arguably improve on that, as it entails that P → P is a tautology, but less intuitively it allows that both P → P and P → ¬P are true for indeterminate P (as a consequence the law of excluded middle and non-contradiction are indeterminate). A non-truth-functional three-valued approach, such as supervaluationism, can also maintain P → P as a tautology, but whilst keeping the law of excluded middle and non-contradiction as true. Hyde (2008: 191-2) argues that the non-truth-functional disjunction means the theory is less appealing (Edgington 1997: 304-5 and Keefe 2000: 96-8 disagree).

Degree theoretic approaches also introduce truth-values in between truth and falsity, but suggest that there are infinitely many of them, represented by the real numbers in the interval [0, 1]. Infinitely many truth values can model the smooth gradient from (say) red to orange in a sorites series. Clear cases of a vague predicate have the relevant property to degree 1, but this degree of truth gradually decreases over a sorites series as the hue becomes less red, and more orange, until we reach the clear orange cases, which are red to degree 0. It might be thought attractive that this sort of approach can distinguish between borderline cases. Given two borderline red/orange patches *a* and *b, a* being noticeably closer to red than *b,* one might be considerably more comfortable calling *a* red than *b,* or still being hesitant to do so, be more confident *a* is red than *b*. Degree approaches can easily model this, by distinguishing in the degree of redness between *a* and *b*, we can model *b* as much closer to 0.5 red than *a*. This isn't so easily done by three-valued logics or supervaluationism, which would give *a* and *b* the same semantic status of indeterminate.

Truth-functional degree theories include Machina (1976) and Smith (2003). Here, the degree of truth of '¬P' equals 1 minus the degree of truth of P. The degree of truth of a conjunction

equals the degree of truth of the least true conjunct (the 'min rule'), the degree of truth of a disjunction equals the degree of truth of the truest disjunct (the 'max rule'). A conditional is always true where the antecedent's truth value is less than or equal to the consequent's truth value, otherwise it is the sum of 1 minus the degree of truth of the antecedent and the degree of truth of the consequent.

Borderline cases are easily modelled, they take truth values between 1 and 0. An apparent lack of sharp boundaries is meant to be captured by the gradient of truth values between truth and falsity. These theories can solve the sorites paradox in different ways. One way is to claim that the inductive premise is close to true, but the argument is invalid. Each instance of $\forall n(\mathrm{R}an \rightarrow \mathrm{R}a(n-1))$ will be close to full truth, assuming the degree of truth of R*a**n* and of R*a**n-1* are close to each other. As the sorites paradox goes from (a least partially) true premises to a wholly false conclusion (i.e. R*a1*), the argument cannot be valid.

There are also non-truth-truth functional degree theoretic accounts. A well-known example is Edgington (1997), which makes use of a notion of 'verities'. Verity comes in degrees, and is meant to represent closeness to clear truth. Her connectives work in similar ways to probabilistic ones, and requires a notion of conditional verity; for instance, the verity of the conjunction of A and B is the product of the verity of A and the conditional verity of B given A (something like what verity B would have were A fully true). The verity of a conditional ($A \rightarrow B$) equals the conditional verity of the consequent on the antecedent. This can give an attractive result, compared to the three-valued logic, for instance. If *a* is a borderline case of redness, and red to degree 0.5 (or '*a* is red' has verity 0.5), 'if *a* is red then *a* is red' will still be true to degree 1, as assuming the antecedent is fully true, the consequence would also be fully true (if *a* is fully red, *a* is fully red). The truth-functional degree accounts would get the same result, but not do so well in other cases. Edgington's account would find 'if *a* red then *a* is not red' fully false, as the verity of '*a* is not red' on '*a* is red' is 0. However, assuming '*a* is red' and '*a* is not red' are both true to 0.5, a truth-functional degree theory would find the conditional fully true, whilst the three-valued semantics would find it indeterminate.

Edgington's theory gives an interesting solution to the sorites paradox. She takes valid arguments to be those that have the *verity constraining property*; that the total unverity of the premises doesn't outweigh the unverity of the conclusion (unverity being equal to 1 minus the verity of the relevant proposition). This means that we can accept the premises of the sorites paradox, it's validity *and* that the conclusion is false. Each instance of the inductive

premise ($\forall n(\mathrm{R}a_{n} \rightarrow \mathrm{R}a_{n-1})$) will be close to truth. If the verity of $\mathrm{R}a_{n-1}$ is slightly lower than that of $\mathrm{R}a_{n}$, then the verity of the conditional will be just slightly lower than 1 (presumably anyway, precisely how conditional verity works isn't entirely clear, see Keefe 2000: 100 for example). This means for any number of pounds, we can find it plausible that if someone were rich with that number, then someone would be rich with just a pound less. But as each of these conditionals will have a very slight unverity, the sum of them all can equal the total unverity of the absurd conclusion of the sorites paradox, unverity 1.

Paraconsistency: Paraconsistent theories don't seem to be as popular at the other non-classical options discussed above, but have received some attention. The most commonly discussed version is *subvaluationism*, the logical dual of supervaluationism. Whilst subvaluationism is defended in Hyde 1997, more recently Hyde has only argued that subvaluationism is at least no worse than supervaluationism (e.g. 2010), but not directly argued in favour of it. In a similar vein, Cobreros (2010) argues subvaluationism fairs better than supervaluationism in some theoretical respects, but doesn't argue directly it is the best theory of vagueness. But given that supervaluationism is regularly considered a promising theory of vagueness, if these arguments are good ones, then it certainly means that subvaluationism is also worth attention.

As mentioned, subvaluationism is the dual of supervaluationism. It also uses the notion of admissible precisifications, but defines truth as truth on *some* admissible precisification (sub-truth), and falsity as falsity on *some* admissible precisification. Propositions that are true on some but not all admissible precisifications will thus come out as both true and false. This is the way the theory captures borderline cases, they are modelled as falling in a truth value *glut*. Allowing gluts, the logic is must be paraconsistent, that is, the law of explosion is invalid. Like other paraconsistent logics, *modus ponens* fails for subvaluationism; if A is sub true, A and $\mathrm{A} \rightarrow \mathrm{B}$ will be true, but B may not be true on any admissible precisification. As well as a non-classical logic, there is a non-truth-functional semantics, the most striking result being that adjunction fails. 'P' and '$\neg$P' can both be true (each being true on some admissible precisification), but 'P & $\neg$P' will always be false, as it can never be true on any admissible precisification (each being a classical valuation).

Subvaluationism can account for the features of vagueness offer a solution to the sorites paradox. As noted, borderline cases are those that are true and false. Particularly interesting is the way subvaluationism can account for tolerance, because of a non-transitive

conditional/bi-conditional. For instance, say A is true on all admissible precisifications, B true on some but not all admissible precisifications, and C is false on all admissible precisifications. Then A → B is true, and B → C is true, but A → C is just false (as on every admissible precisification A is true and C is false). The failure of this classical inference means we can accept for every pairwise cases in a sorites series that if one is rich (say) then so is the adjacent member, but needn't accept that if the first member is rich then so is the last member. This gives one sort of solution to a type of sorites paradox, which finds it invalid. To solve the sorites paradox in the form we've been considering it, with the universally generalised inductive premise, subvaluationism can employ the same method as supervaluationism and degree supervaluationism; the inductive premise is false, as false on all admissible precisifications, and thus the paradox is unsound.

An alternative to subvaluationism is the stronger paraconsistent logic *LP*, first presented by Priest (1979), and suggested as the basis for a theory of vagueness by Ripley (2012) and Webber et al (2014). LP allows that contradictions may be true, i.e. a proposition and its negation, and accordingly denies the validity of explosion: it is not the case that any proposition is entailed by a contradiction. I say LP is a stronger paraconsistent approach because unlike subvaluationism, it entails that propositions of the form 'P & ¬P' *can* be true. Ripley suggests that this is an improvement on subvaluationism, claiming it's more intuitive that the conjunction of any two true propositions is also true (2012: 336), and prefers LP for this reason.

This theory can also explain the phenomena of vagueness. Borderline cases come out as those that are true contradictions, both true and false, and it follows from LP that the sorites paradox is *invalid;* modus ponens fails for the standard material conditional. Furthermore, LP can give the same account of tolerance as subvaluationism through a lack of transitivity in the conditional. We can claim that for every pairwise member of a sorites series, one is an instance of the relevant predicate if the other is, without be committed to saying that if the first member of the series is an instance of the relevant predicate, so is the last member (as discussed in detail in Webber et al 2014).

Epistemicism: The theories we've considered so far have placed the focus on new logical or semantic treatments of vagueness. Epistemicism (Williamson 1994) keeps logic and semantics classical, and claims that vagueness is just a matter of *ignorance*. The suggestion is that all 'vague' language is in fact precise, there are sharp boundaries in the extension of

every apparently vague word. This means there is a number of pounds that would make one rich, but a pound less would make one not rich, there is a shortest height one can be and remain tall, a lowest number of words a thesis can contain and remain long. Accordingly, all vague language does in fact make sharp divisions, and everyone is either rich or not rich, tall or not tall, old or not old. This gives a straightforward solution to the sorites paradox; the inductive premise is false, there is a number of pounds that one can possess and be rich, but would not be rich with a pound less.

The location of these sharp boundaries, which supervene in an unknown way on language use, are unknowable, due to margin of error principles. Williamson claims that if a belief is to count as knowledge, then it cannot be down to luck that it's true. If you believe P, which happens to be true, but would also believe P in a very similar situation where it is in fact false, then you do not know that P. Say having read this thesis, you decided to make guess at how long it is, and take up the belief that it is 70,594 words long, and it turns out that this belief is *true*, is it knowledge? Williamson would argue not; had the thesis been a single word shorter, you would not have noticed the difference, and would have taken up the same belief. In this respect, your belief was lucky. For the same reason, speakers cannot have knowledge about where the boundary in a vague word's extension is. If I take up the belief that £500,000 is the least one can possess and still be considered rich, based on what I know of the competent usage of 'rich', and I am in fact right, this would not be knowledge, as in a very similar situation where usage actually placed the boundary at £500,001, I would not have been sensitive to the difference. Thus, any beliefs about the whereabouts of these boundaries could not count as knowledge. Vague language gains unknowable sharp boundaries through usage, in an unknown way. Borderline cases are those cases which are close to the boundary, and as we don't know where the boundary exactly is, we don't know which side of the boundary they fall.

The theory seems to many simply too incredible to be true: it just seems wrong that there is a precise number of seconds that is the most that can pass in a person's lifetime while they are young, and second more is sufficient for them to be 'not young', that a rich person could become not rich by losing a single pound. But this is of course far from a knock down argument, and indeed knock down arguments against epistemicism are hard to come by (as they are for all theories of vagueness). As far-fetched as it may seem, epistemicists can simply bite the bullet on this, and appeal to the attractive parts of their theory, namely the full preservation of classical logic and semantics.

2. Conclusion

This chapter should have clarified what vagueness is, and the goals of theories of vagueness that set out to model it. As is clear from this brief overview, there is no shortage of possible directions to go on vagueness, the literature has provided plenty of competing accounts. How are we to decide between them? As mentioned, knock down arguments aren't usually available when it comes to theories of vagueness. Each theory comes with costs and benefits, and people have to weigh these up, arguing that on the whole a particular theory does the best job of modelling vague language use, retaining intuitions we have whilst limiting counter-intuitive consequences.

In the following chapter, I will be arguing that experimental philosophy can have a theoretically interesting role to play here. I argue that the primary goal of the theories of vagueness is to provide a descriptive theory of how vague language works, and accordingly empirical data about language use can help create a clearer picture to model.

Chapter 2 The Experimental Approach to Vagueness

The aim of this chapter is to develop and defend the role of empirical data in theorising about vagueness. Interest in experimental philosophy has been growing in recent years, and philosophers have employed the findings of studies to support arguments in debates surrounding epistemology, ethics, free will, philosophy of language, logic, amongst others. There is plenty of discussion of the use of empirical methods in philosophy, including plenty of criticism. However, there are lots of different types of project that come under the umbrella of 'experimental philosophy', and not every criticism can be appropriately directed at every sort of approach. The first task here is to distinguish the role for empirical data when it comes to vagueness from other types of project in experimental philosophy. Some see experimental philosophy as a *psychological* investigation, a tool to learn how intuitions operate and how the human mind works. Others use it as a tool in a *negative* project, arguing that instability and variation of intuitions undermines the use of them in the methodology of 'armchair philosophy'. Empirical work in the area of vagueness is most usefully seen in the light of what has been called the *positive* project; its aim to deliver useful information that can develop and inform philosophical theorising.

Seen in this way, there is good reason to think experimental work is relevant to philosophical research on vagueness. My argument here will rest on two main claims: first, that empirical work is relevant to theorising about vagueness *if* we're engaged in a *descriptive project*, and second, that any interesting work on vagueness *is* essentially descriptive. With the first premise I am largely endorsing the view of David Ripley; if the aim of philosophical theorising about vagueness is to model a real-life phenomenon, how vague expressions and concepts *do* work, then experiments are (at least theoretically) a good way of getting a clearer picture of the phenomenon to model.

However, I will differ from Ripley in an important way. Ripley leaves the door open to the possibility that some philosophers might not be engaging in this descriptive project, but rather a prescriptive one; i.e. one that aims not to capture how speakers *do* use vague language, but rather *should*. I think we should keep this door mostly closed. The second premise of my argument claims that any interesting work on vagueness aims to model the logic and semantics of the actual concepts we employ. I think most philosophers take themselves to be

discussing the nature of our real natural language, and this is clear from the nature of the debate. Furthermore, I'll argue any theory which aims to be purely prescriptive would be uninteresting.

I take my argument to successfully show that experimental work can (in theory) inform theorising about vagueness. I'm aware there is a lot of scepticism about the experimental approach in general, which should be addressed. There are lots of criticisms in the literature, but not all of it can be appropriately aimed at the approach I defend here. For example, there is significant discussion of the *negative* project[1], and the role of intuitions in so called armchair philosophy, but this isn't relevant here. I will finish the chapter with a reply to an objection that is relevant; the expertise objection.

1. Different Experimental Approaches

As we saw in the previous chapter, there are several theories of vagueness to choose from, each claiming to give the semantics and logic of vague discourse, account for the apparent features of vague language and solve the sorites paradox. Each of these have been carefully developed and argued for by philosophers over a long period, academics who are experts in the philosophy of language and logic, and have spent a great deal of time thinking about these issues. Given this, it is understandable that there'd be initial resistance to the suggestion that the right way to approach the problem of vagueness is by asking the general public, the 'folk', about it. This approach takes the focus away from the arguments of the philosophers who have developed and defended these theories, and places more weight on information about the linguistic practice of normal speakers. Understandably, philosophers may be suspicious of this, but there is a good case to be made in favour of it. Before speaking in direct defence of it, I will outline a few way philosophers have used experimental tools, to better clarify what I mean by experimental approach here.

'Experimental philosophy' can be conceived of in different ways, but in general involved gathering and using empirical data to support philosophical conclusions. This most often involves questionnaires and self-report type studies, which usually involves asking subjects to judge the semantic status some claim (e.g. ''Gödel' refers to the person who discovered the incompleteness of arithmetic'), but studies can include other measures, such as eye

[1] This distinction is not my own and is quite widely acknowledged in the literature, see Hansen (2015) for example

movements and reaction times (see Krifka 2011 for instance). There are different possible approaches within the experimental project, with different philosophical goals in pursuit of which to employ empirical data. I will quickly outline a couple which are not the sort of thing I am recommending here.

Perhaps the most widespread, and widely discussed, approach is what has been called the *negative* project. This refers to studies which seek to show that the methods used by philosophers in their arguments rely too heavily on intuitions. This view is summed up by the following quote from Swain, Alexander, and Weinberg:

> [There is] an existing body of empirical research demonstrating that intuitions vary according to factors irrelevant to the [philosophical] issues. …[This research] raises questions about the reliance on intuitions…. We take the growing body of empirical data impugning various intuitions to present a real challenge for philosophers who wish to rely on intuitions as evidence. (2008, 153)

Plenty of research has been conducted with this kind of goal in mind. A paradigm influential example is a study by Weinberg, Nichols, and Stich, who hypothesise that epistemic intuitions vary from culture to culture (2001). They investigate participant responses to 'Gettier cases', cases which were famously used to argue that in certain cases justified true belief is not sufficient for knowledge (see Gettier 1963). In their study, Weinberg et al find that the intuition that such cases aren't sufficient for knowledge isn't consistent across cultures, with East Asians more likely to report it is knowledge than Westerners. Weinberg et al conclude that as the intuitions relied upon in western philosophy isn't shared across cultures 'it is hard to see why we should think that these intuitions tell us anything at all… about epistemic norms or indeed anything of philosophical interest' (pp. 452). There are many empirical studies that make this sort of argument. Petrinovich and O'Neill (1996) show that moral intuitions about trolley problems can vary based on the wording of examples, and conclude that moral intuitions cannot track the truth, Knobe (2003) finds intuition about intentional action can be also manipulated based on wording, arguing it casts doubt on intuitions used in the philosophy of action, Machery et al. (2012) find variation in intuitions about the reference of proper names, and argues it undermines the reliance on intuitions of philosophers as evidence of what the correct theory of reference is.

The *negative* project is not the sort I will be arguing for here in relation to vagueness. I don't have in mind that we can use experimental philosophy to undermine the use of intuitions in philosophical argument, but rather something much more constructive. Another sort of project that can come under experimental philosophy umbrella, is a *psychological* one. This

is the kind of approach argued for by Knobe and Nichols in their *Experimental Philosophy Manifesto* (2008), as seen in the following quote:

> It used to be a commonplace that the discipline of philosophy was deeply concerned with questions about the human condition. Philosophers thought about human beings and how their minds worked… The new movement of experimental philosophy seeks a return to this traditional vision. Like philosophers of centuries past, we are concerned with questions about how human beings actually happen to be. (pp. 3)

They see three different strands of this project and its usefulness to philosophy. One, '*sources and warrant',* aims to learn more about why people (importantly, philosophers) have the intuitions they do, and thus assess if an intuition is a good source on which to form a belief. If we know more about why a philosopher reflects and intuits that an agent *is* morally responsible in some situation (for example), we can determinate if their intuition should be trusted in this case or not. Some studies have been carried out with this approach in mind, for instance Greene et al (2001), who use evidence from an fMRI to show why participants tend to respond in particular ways to certain moral cases, and Nichols and Knobe (2007) investigating the role of 'affect' (emotional responses) on intuitions relevant to the compatibilist/incompatibilist debate.

A second, '*diversity',* connects with the negative project discussed above, and focusses on cross-cultural differences in intuitions. Their intention isn't to investigate differences in intuitions to challenge philosophical methodology however, but to learn more about the origins of our intuitions:

> If I find out that my philosophical intuitions are a product of my cultural upbringing, then, since it's in some sense an accident that I had the cultural upbringing that I did, I am forced to wonder whether my intuitions are superior at tracking the nature of the world, the mind, and the good. These are manifestly philosophical questions. And to determine the answers, we need to know a great deal more about both our own intuitions and those of other cultures. (pp. 10)

A third, '*mind and its workings',* simply aims to better understand humans and the workings of their mind:

> The philosopher on one end of the hall can be developing complex mathematical theories about the relevance of Bayesian inference to causal modeling, while the philosopher at the other end of the hall can be developing complex theories about how people's causal intuitions reveal some fundamental truth about human nature. (pp. 11)

Unlike '*sources and warrant',* which investigates psychological mechanisms to help philosophical inquiry (by helping to identify if an intuition is a reliable source or not), this project takes interest in psychology in of itself. As seen in an earlier quote, Knobe and

Nichols welcome the inclusion of psychological interest in philosophy; they see it as a move back to a more traditional view in philosophy, that of ancient and early modern philosophers.

This psychological project is also not the sort I have in mind here. Experimental philosophical work in vagueness can most usefully be done in what might be called the *positive project.* Alexander et al outline a broad view of this in the following way:

> [This program] concern[s] traditional analytic philosophy's practice of appealing to philosophical intuitions as either evidence for (or against) philosophical claims or data both about the nature of our folk philosophical concepts and judgments and about the nature of the domains in which we make those judgments… According to experimental philosophy's "positive program", experimental philosophy is (at least an indispensable part of) the proper methodology for this practice. (2010: 297)

According to the *positive* project, experimental work can be part of a philosophical investigation into some debate, issue or phenomena. Not in the sense of restricting the philosophical method, or taking an interest people's cognitive processes, but to help shed light on philosophical problems. Many examples of this type of experimental philosophy occur in the field of the Philosophy of language. This is because the broad aims of the area involve understanding the workings of natural language; e.g. issues around meaning, reference, truth, formal semantics. That the aim of philosophical investigation in this area is to understand a natural phenomenon better (that of natural language), it isn't surprising that empirical work on how language does work might be helpful (this is ultimately why I claim vagueness is amenable to empirical investigation).

There are plenty examples of this. For instance, some studies have investigated the meaning of *natural kind* terms, words that refer to a naturally occurring group such as 'tiger', 'lemon' and 'gold'. A traditional view (e.g. Mill 1884) has it that the meaning of such terms is a set of properties, for example 'tiger' might mean 'large carnivorous quadrepedal feline, tawny yellow in colour with blackish transverse stripes and white belly' (Kripke 1980: 119). Kripke and Putnam (1975) disagree with this, and favour an alternative, *essentialism,* according to which a natural kind term refers to the 'essence' of the relevant kind (Kripke 1980: 136-8).

According to the traditional view, the meaning of a natural kind term is partially fixed by speaker's beliefs about that kind (the meaning of 'tiger' is partially fixed by speaker's beliefs that they have 4 legs etc…), whereas on the essentialism, 'tiger' can refer to the natural kind regardless of speaker's beliefs about them (even speaker's beliefs about tigers were all false for example).

Some studies have sought to test these theories empirically. Braisby et al (1996) suggest that essentialism predicts reference of natural kind terms is determined independently of speaker's beliefs about the relevant kinds, and that this will manifest in linguistic behaviour. They test a thought experiment featured in Putnam (1975), where it is discovered 'cats' aren't actually mammals, but robots controlled by Martians. Putnam, and the essentialist view, predicts that 'cat' still refers to the robots despite speaker's false beliefs about them. Braisby et al find that in one study 58% of subjects share this intuition, and 76% do in another study. Braisby et al take this to reflect a divergence from essentialism, as a significant number of subjects held the intuition that 'cat' doesn't refer to the robots controlled from Mars.

More interestingly, Braisby et al find subjects assenting to apparently contradictory pairs of sentences:

1. Tibby is a cat, though we were wrong about her being a mammal
2. Tibby is not a cat, though she is a robot controlled from Mars

(1) reflects the essentialist intuition, and (2) reflects an intuition in line with the traditional view. It is striking that participants would agree to both of these sentences, as they disagree about the reference of 'cat'. The authors use this to promote a 'representational change' theory, according to which 'cat' and other natural kinds can be ambiguous in their reference; they can refer in line with both the essentialist and traditional views. Speakers assenting to both (1) and (2) above is taken as evidence that they can switch between interpretations: in some contexts, they take 'cat' to refer *essentially*, and in others *traditionally*. Jylkka et al (2009) investigate the same debate, and take themselves to find evidence in favour of essentialism, however Hansen (2015: 16) interprets their results as also supporting the 'representational change' theory (and provides a good overview of this whole debate).

This is an example of an investigation in line with the *positive* project in experimental philosophy. Here philosophical theories about the meaning of natural kind terms have been investigated empirically based on intuitions of speakers, delivering what is arguably positive evidence for a new theory of their meaning. There are many other examples within the philosophy of language; including investigations into the semantics of colour and scalar adjectives (Hansen and Chemla 2015), how context affects the extension of 'knows' (Buckwalter, 2010), the semantics of 'most' (Pietroski et al 2009), and epistemic modals (Knobe and Yalcin, 2014).

This is how I see the role of experimental work on vagueness, as providing information that can inform philosophical theorising. As we'll see in the following chapter there are already several studies that have been conducted on vagueness in this vein, that could be added to the above list. Before getting to that, I will discuss in detail why I think this approach is an appropriate method of study.

2. In Defence of the Experimental Approach

I now turn to discuss why I think experiments can be a genuinely useful tool to aid theorising about vagueness. My main argument is based on two simple claims:

1. If the goal of theorising about vagueness is to develop an accurate descriptive theory, then experimental work is relevant to this theorising
2. The goal of theorising about vagueness is to develop an accurate descriptive theory

I'll argue for these in turn. Premise 1 is quite plausible; it seems that *if* we're interested in drawing an accurate picture of a real-life phenomenon, that of how our vague language works, then data that informs us of what the phenomenon is like is relevant; it gives us a clearer view of what we're trying to model. Premise 2 may be less obvious, but reflecting on the nature of our theorising to date, that is the goals we set for our theories, the way philosophers weigh up theories, and the arguments philosophers set out for and against theories, reveals that theorists have been engaged in an importantly *descriptive* project. Furthermore, any project that was wholly prescriptive would be very uninteresting. We could easily get a semantics and logic for a set of concepts that is free of paradox for instance, as an alternative to the set of concepts we actually use, or give an alternative way of conducting our concepts that avoids paradox, but that isn't what we were ever interested in; we wanted to know how our actual apparently paradoxical vague concepts operate.

The first question is to clear up what I mean by 'descriptive' and 'prescriptive'. I have in mind something like the distinction Dave Ripley draws:

> There are at least two questions in the area, though, that are more likely to generate conflict between distinct hypotheses: first, how *do* we manage to draw distinctions using vague talk and thought? and second, how *should* we do it? Following Scharp (2013) (on a different topic), call these the *descriptive* and *prescriptive* questions, respectively. (2016: 525)

A descriptive theory aims to answer the 'how *do…'* question. The goal of philosophical inquiry about vagueness is to give the details of *real-life* language use; a model which tells us

how vague language operates, what its logics and semantics are, explains borderline cases etc… A prescriptive theory on the other hand aims to answer the 'how *should…'* question. The goal of philosophical inquiry in this case is not to model features of the vague language we actually use, but something else. It's not entirely clear what that something else is. Ripley doesn't offer any further elucidation of the distinction between prescriptive and descriptive theories of vagueness, and as I'm about to argue that philosophers are interested in a descriptive project, I don't have a clear proponent of 'prescriptivism' to offer a good outline of the project.

The distinction could be similar to that between prescriptivist and descriptive linguistics (see e.g. Brown, Attardo and Vigliotti 2014: chp.1). A descriptive linguistics seeks to understand the rules that do govern language, whereas a prescriptive linguistics sets rules that come from outside of the language (perhaps based on rules used at earlier stage). For instance prescriptive linguists have claimed that people oughtn't used 'split infinitives', that is, have a word in between 'to' and the rest of an infinite. On this view, 'to boldly go where no man has gone before' isn't grammatical. In contrast, a descriptive linguist would be more interested in actual language usage, and accounting for the rules that govern that, which wouldn't necessarily rule out split infinitives (given their common usage). It should be noted that the descriptive theory still takes language to be *rule governed*, and thus there is still a right and wrong way to use language: successful language use requires following those rules. In this way, the theory still makes normative claims of a sort, i.e. in giving conditions for successful language use. What is important is that these rules are based on the linguistic conventions of real life speakers, and that is what the theory aims to capture, in contrast to a prescriptivist linguistics that imposes rules without interest in how natural language is actually used (see e.g. Kim and Sells 2008: 4).

In a similar vein, here I think of descriptive theories of vagueness as those that try to capture features of actual natural language. But by modelling vague language, they are giving a theory of *correct* language use, which is normative in a sense. There is a correct and an incorrect way to use natural language, and theories of vagueness capture part of this correct or competent usage. However, this correct usage is determined by our naturally occurring linguistic conventions. Just as in the case of descriptive linguistics, when we have a theory of correct usage, whilst that is in a sense a normative theory, it is ultimately a descriptive one, as we get a correct usage by a description of the linguistic conventions. A prescriptive one wouldn't be interested in accounting for the semantics of actual natural language use, but of

giving theory of how we *ought* make distinctions using language, regardless of how speakers actually do it.

I take it the sense in which a descriptive theory of vagueness is normative in being a theory based in *competent* language use is clearly distinct from the type of theory I am thinking of as primarily prescriptive. This is a theory which pays no attention to linguistic convention, but rather aims to give an *alternative* to natural language which is paradox free. I cannot give many clear examples of this from the literature, as I do not think philosophers are trying to do this for the most part. Someone pursuing this kind of theory might follow the suggestion of Burgess and Plunkett, in the research program they call *conceptual ethics*, which (very broadly) pertains to 'the nature or methodology of semantic and conceptual *prescriptions*' (2013: 1091). Burgess and Plunkett note that semantics and conceptual analysis tends to be a descriptive project, concentrated on how speakers/thinkers actually employ the 'meanings or contents of representational vehicles', but they see a possible role for a normative project too, which involved refining or replacing our everyday concepts with new ones. For instance, on the topic of the inconsistency and truth (referring to the liar and sorites paradoxes), they write:

> '…those of us loath to accept contradictions might naturally wonder whether we ought to revise classical logic or our ordinary notion of truth in light of inconsistency theory. On the one hand, the paradoxes obviously pose no purely practical problems in everyday conversation or deliberation. But TRUTH and logical notions also figure in theoretical enterprises like mathematics and linguistics. Perhaps we should introduce consistent surrogates for our defective expressions in these domains, lifting candidates from the technical literature on the paradoxes'. (pp. 1093)

A philosopher, in pursuit of the prescriptive 'how *should* we do it' question, might identify our language as defective, and suggest an alternative improved way of conducting language. I think the closest approach to this in the literature on vagueness is the one attributed to Frege and Russell. Frege saw vagueness as a defect of natural languages, along with ambiguity, and was instead committed to developing artificial languages which could deliver the precision needed to philosophy and science. In the *Grundgesetze*, for instance, he says that concepts must have sharp boundaries, and cannot tolerate objects that may or may not fall in their extension; he says that a region of application of a concept with blurred boundaries 'would not really be a region at all' (1903: §65). Similarly, in the *Begriffsschrift,* Frege uses the sorites paradox to show why vague predicates shouldn't be admitted into the language (1879: §27). I take it that this is the kind of thing a prescriptive theory would look like; a logic and

semantics for a language which free us of the 'defects' of the expressions of natural language, a new set of concepts to supersede those speakers use.

2i. Premise 1- empirical results are relevant to descriptive philosophical theorising

I suggest, following Ripley (2016), that the descriptive project is amenable to empirical investigation. If we want to discover and model how speakers use vague language, then information about competent language usage will surely be useful. As Ripley says: 'to find out how we in fact *do* get a job done, armchair reflection can only go so far' (256). Experiments can play a useful role alongside arm chair reflection, providing information about vague language use, and giving a clearer picture of the phenomenon to model.

Ripley uses an example to show how empirical data might be relevant to a descriptive theory of vagueness (2016: 526). The general idea is that if our theories are descriptive, then we might expect them to make some sort of predictions about how speakers use vague language, or mentally represent vague concepts, predictions which can be then tested as hypotheses. Ripley takes *epistemicism* as an example. As discussed in the previous chapter, epistemicism takes it that vagueness is a matter of ignorance; all vague language does draw sharp distinctions, its vagueness consists in our ignorance of the whereabouts of the boundaries it draws. If this is right, then we might expect it to show up in the way speakers deal with vague language. For example, Bonini et al. (1999), who seek to support epistemicism with experimental evidence, claim epistemicism makes a prediction like the following: 'a typical speaker *S* of a natural language... mentally represents vague predicates in the same way as other predicates with sharp true/false boundaries of whose location *S* is uncertain.' (pp. 387). If epistemicism is right that our vague language is actually precise, but we do not know the location of the of the sharp boundaries it draws, then we might think we would handle vague language and concepts in the same way some others which are also precise, but not vague, and whose sharp boundaries we do not know the location of. For example, in the study, they investigate whether 'tall' is treated in the same way as 'of average height among 30 year old Italians'. It is *suggested* that the latter predicate is precise (although perhaps there is reason to be unsure that it is; there may be vagueness about who counts as 'Italian' for instance, or ambiguity about 'average'), but that subjects in the study will not know what height is the average for 30 year old Italians; it draws an unknown sharp boundary. According to the above hypothesis, the epistemicist should predict that the two predicates will be mentally represented in similar ways.

Ripley claims that 'it should be clear that this hypothesis is both intimately related to epistemic theories of vagueness and open to experimental study' (2016: 256). There are of course some complications though. As Ripley points out, epistemicism doesn't have to predict the above hypothesis; an epistemicist might hold that not only are we ignorant of the location of the sharp boundary, but also ignorant that vague extensions are sharply bounded at all. If this were the case, then speakers might represent vague language differently to other precise language with unknown sharp boundaries which the speaker knows is precise. Ripley suggests that this too can be tested experimentally, it's just a different hypothesis. A more pressing worry is the differences between vague predicates and the 'other predicates with sharp true/false boundaries of whose location *S* is uncertain' on the epistemicist picture. According to epistemicism, the location of sharp boundaries of vague language are *unknowable*, on account of margin for error principles (see Williams 1994: chp. 8). But this is not the case for a predicate like 'above average height'; with enough research we could determine the sharp threshold to pass average height, but we couldn't find the sharp threshold for 'tall'. This raises reasonable worries about how similar the mental representation of the two should really be expected to be.

Whilst this is an interesting thought, it's a worry for the set-up of this particular study, not for the general experimental approach they are taking, which is what I'm interested in for now. Regardless of the plausibility of the hypothesis Bonini et al set out (in chapter 3 I'll discuss Bonini et al in detail), the general approach here looks promising. If we're interested in the descriptive project, we want our theories to give an answer to the 'how *do'* question, i.e. provide accurate models of the semantics and logic of vague language. If theories are trying to tell us the features of our actual vague language, then we can reasonably expect what these theories say to show up in linguistic practise and the mental representation of vague concepts, which in turn can give us testable hypotheses.

In general, there are clearly experimental results which could affect the plausibility of a (descriptive) theory of vagueness. For instance, say that a study is run with a significant sample size, and we find that subjects always reliably assent to contradictory descriptions of borderline cases. When shown a borderline red/orange colour patch, the subjects assent to 'definitely red and definitely not red', they assented to 'definitely tall and definitely not tall' for a borderline tall person, 'definitely rich and definitely not rich' for someone borderline rich, and so on. Assume also they reliably didn't assent to what might be thought of as a corresponding 'gappy' description; 'neither definitely red nor definitely not red' etc. It is

clear that such a strong empirical result would have an impact on theorising about vagueness. It would add weight to some paraconsistent theories. A model which allowed true contradictions (e.g. Webber et al 2014), would look like it was accurately modelling vagueness based on this data; if speaker's use of language reliably showed borderline cases being described as true contradictions, a model which allowed 'P & ¬P' to be true would be attractive. Similarly, other types of theories may look less attractive, without a substantial explanation of the result. For instance, according to supervaluationism, contradictions are always false, and the law of excluded middle always true. Where it has been found that speakers reliably assent to the above contradictory descriptions of borderline cases, then supervaluationism would look less attractive as a theory of vagueness; we would need a good story about *why* speakers reliably assented to the contradictions, given they are always false. I imagine that in such a scenario, a paraconsistent model that can simply say speakers assent because the contradictions are true, would look like a more promising theory of vagueness.

I take this to show that there is the potential for a role for experimental work in descriptive theorising about vagueness. As argued by Ripley, we can reasonably expect the claims of descriptive theories of vagueness to manifest in the way speakers deal with vague language, and that this in turn can give us testable hypotheses. The above imagined experimental result paints a plausible picture of how data could affect theorising in this area. Of course, the result is very idealised, but that needn't take away from the result that there is a role for experimental data (in theory) here; in reality results will likely not be so clear cut, but that needn't mean that relevant results haven't been found. The point is that there *is* the potential for empirical work to play a role in theorising about vagueness where we are engaged with the descriptive project; a real-life empirical result will likely have messier data, but it will simply take more work to draw out the philosophical significance in those cases.

2ii. Premise 2- the descriptive project is the main project

At this stage, one could accept that where philosophers are trying to answer the 'how do we do it' question, there is a role for experimental data, and that alone is something of a success for the experimental approach. Sure, experiments won't be important for those answering the 'how should we do it' question, but they weren't mean to be; as long as they are useful for the descriptive question, they are useful.

This is where Ripley leaves his argument. I want to go slightly further than this, and argue that the descriptive project is the one philosophers are rightly primarily interested in; this is what I claim in the second premise of my argument. I think it's quite clear that philosophers are interested in a theory of how our *actual* vague language works, based on the nature of the debate, and the arguments put forward.

To start with, consider the experimental result I describe above, which finds that subjects reliably describe a borderline red/orange colour patch as 'red and not red'. If it such a robust result were found, I suspect that it would prove decisive in the current debate. If it were found that every subject reliably assented to 'red and not red', and the relevant contradictions for other vague concepts, to describe borderline cases, then many would surely agree that paraconsistency was a good theory of vagueness. Or if they wished to still defend an alternative, would have to explain this result, and how their theory remains consistent with it. That there are conceivable empirical findings so strong and robust that the plausibility of theories would be affected in the contemporary debate reveals that the project they're discussing is descriptive, to some degree at least. Here I may be begging the question against the philosopher who doubts experimental work, but I do suspect that the result I described would receive significant attention, and rightly so.

It is quite clear from the nature of the discussion around vagueness that philosophers are engaging with a descriptive question. Some philosophers are quite explicit that this is what they have in mind. For example:

> It is not enough to know that 'heap' is vague; in what respects is it vague, and to what degree? An answer to the question is part of a description of the ordinary uses of the word 'heap'. (Williamson, 1994: 71)
>
> We should first determine how vague expressions function, and figure out the logic from there. (Shapiro 2006: Sect. 4.5)
>
> This semantic but classical approach comes into view when we ground our theorising about vagueness as deeply as possible in common-sense and our actual competent use of vague words. (Raffman 2014: 12)

Here the authors show that the nature of our natural vague language is important to their theorising, it is this they aim to model. The volume *Vagueness and Language Use* (Egré and Klinedinst 2011) gives a set of essays where it is explicit the authors are interested in natural language (in particular at least the relations between vagueness, gradability, comparatives, intensifiers and modifiers). For instance, one essay begins 'This chapter is about vagueness in natural language semantics' (Sauerland and Stateva 2011: 121). Beyond this, the discussion

shows philosophers are interested in our real language use. As Keefe mentions, the methodology of debate surrounding vagueness often takes the shape of a 'reflective equilibrium' (2000: 38). There are rarely knock-down arguments against theories of vagueness. Instead, assessing a theory involves a weighing-up of the desirable intuitions it maintains against the counter-intuitive consequences it comes with. Central to this is the idea that there are a number of important intuitions that we want our theories to maintain. These are some of the things I discussed in the previous chapter. To name some, we might want our theories to allow speakers to make genuine distinctions with vague language, and continue to use it successfully, that speakers will stop applying a concept at some point during a sorites series, that there appear to be no sharp cut-offs to a vague word's extension, that vague language permits borderline cases, that vague language is tolerant, that the sorites paradox is unsound. These intuitions are, it seems to me, intuitions about how our actual language works, which theories try to capture or explain.

I pointed out the way in which theories do this in the last chapter. Supervaluationism, for instance, (arguably) gives a plausible account of borderline cases and the lack of sharp boundaries based on semantic indeterminacy and admissible precisifications. Epistemicism and degree theories also do this, the former based on ignorance, and the latter based on degrees of truth. This is seen as something desirable about the theories. We want them to explain what is going on with vagueness related phenomena, to give explanations of things like borderline cases, or why we don't fall for the sorites paradox. If a theory failed to offer any explanation of borderline cases, apparent lack of cut offs etc… they would be criticised; there'd be part of the phenomenon they aren't accounting for. Were these theories aimed at the 'how should' question, it's not clear what the motivation would be to explain things like borderline cases, or the apparent lack of sharp boundaries; one could just construct a theory where there are no borderline cases, for example. Borderline cases are just a product of how we *do* use language, if philosophers were giving a theory of how we *should* use language, there'd be no need to explain them.

Of course, there can be disagreement about which intuitions are the crucial ones a theory should aim to maintain, as Machina suggests, debate about them can come down to just 'a battle of raw intuitions' (1976: 51). In this set up we can see a good role for experimental work; helping to determine what the important intuitions theories should maintain are. Keefe writes:

> …there can be disputes about what is in the relevant body of opinions- whether some given opinion is really one that we must attempt to save. It may take a (carefully formulated) questionnaire to discover what the opinions of the folk really are. (And it must not be assumed that the corrupted views of the theorising philosopher reflect the common view). (2000: 41)

This reinforces both the suggestion that the project is descriptive, and that there is a role for experimental philosophy. Here Keefe agrees that the body of opinions to be explained by a theory of vagueness is the one determined by the folk, the speakers of the language, and further that we shouldn't expect philosophers to always be the best judge of what those are. This is where the role of the experimental comes in (such as a questionnaire), to inform philosophers about what the critical opinions that theories of vagueness need to aim to model.

Just as there are a set of intuitions theories aim to capture, theories are criticised for having counter-intuitive consequences, and I think this regularly means, for failing to accurately model a part of the phenomenon of vagueness. There are several examples of this. For instance, consider the following criticisms of epistemicism:

> What then draws the sharp line determining the well-defined extension of 'tall' to which the epistemic theorist is committed? …The epistemic theorist needs to explain away the intuition that *nothing* does. (Keefe 2000: 77)

> …imagine someone, Roger, who thinks his life will go better if the number he chooses for his bank account password ends in the same digit as the last digit of the [number of nanoseconds which is the oldest one can be without being 'old'] … [Roger's belief] seems based on a kind of conceptual confusion about vagueness. (Field 2010, in Dietz and Moruzzi 2010: 203)

> Classical semantics, then, recommends that we should endorse conjunctions of independent borderline propositions much less strongly than we endorse the conjuncts individually. But, as Schiffer observes, this just seems wrong (Schiffer 2003: 204). It seems perfectly appropriate to endorse the conjunctive proposition that Jim is tall and bald and smart to about the same (middling) degree as we endorse the conjuncts separately… classical semantics is not correct for vague discourse. (MacFarlane 2010, in Dietz and Moruzzi 2010: 444)

Keefe argues that there is nothing about language or the world that could set the sharp boundaries postulated by epistemicism, asking here how one could be in theory fixed. This criticism would surely be misplaced if the debate were centring around a *prescriptive* project. It is arguing that nothing about the way we use language *does* set sharp boundaries, and it only makes sense to criticise epistemicism in this way if it is making a descriptive claim; the claim that something does really set sharp boundaries. If we take the theory to prescriptive, the objection would lose force. If it were making a claim like 'we ought to use language as if there are sharp boundaries', then it doesn't matter whether language use and the world could draw sharp boundaries or not; the theory isn't trying to capture the semantics of natural language.

The other two objections also only make sense if we take epistemicism to be a descriptive theory. Field's concern is also aimed at the claim that there really are sharp boundaries, suggesting it is a conceptual confusion. MacFarlane's objection is aimed at the consequence of accepting classical semantics in our theory of vagueness, arguing it doesn't give the correct model of vague discourse. Epistemicism would require a probabilistic reading of conjunctions, but MacFarlane (and Schiffer) think that this isn't the right result, that people would have as much confidence in the conjuncts separately. If the classical semantics postulated by epistemicism is prescriptive, then this objection would slightly miss the mark; why care about how vague discourse actually works, we're telling you how it should work. It is that epistemicism is meant to capture how things actually *do* work that this objection is appropriate.

This pattern is not peculiar to epistemicism, I think many objections to theories of vagueness can be thought of as aimed at a descriptive failing. Supervaluationism has the arguably strange consequence that an existential generalisation can be true, without a verifying instance (any particular object of which it is true). But some argue that this is just not what 'there is' means, and thus suggest that supervaluationism gets the semantics wrong for the existential quantifier (e.g. Williamson 1994: 153-4). Similar comments can be made about the supervaluational semantics for disjunction, for example Hyde argues that the theory doesn't capture the right meaning of the connective (2008: chp 4). These objections only have force if we take supervaluationism to be trying to give the truth-conditions of natural language, rather than giving a prescriptive theory of connectives with alternate semantics that we should use. Degree theories are sometimes criticised for a potential commitment to an infinite number of sharp boundaries (e.g. Keefe 2000: 113). The concern is that by introducing perhaps an infinite number of truth values, we will be presented with the issue of sharp boundaries between them; at what number of hairs does a borderline bald man go from 0.501 to 0.5 bald? The concern is that language use doesn't seem to allow for sharp boundaries at all, let alone infinitely many of them. But again, this objection is only useful if we take ourselves to be engaged in a descriptive project, with degree theories giving one possible model of how vague language actually *does* work, not a theory of how it *should* work. If it were the latter, the objection would lose force; it doesn't matter that our vague concepts do lack infinitely many sharp boundaries, we're giving a theory of how our concepts should work, not how they do.

These objections only have force if the debate is about a descriptive project, about how successful theories are in modelling vague language, the way it *actually* is. That theories aim to account for important phenomena related to vagueness and are criticised for failing to accurately capture intuitions about how our vague language and concepts work, shows that the debate is centred around the descriptive question of vagueness. So, we can see that not only is the case that empirical data can inform theorising, at least insofar as that theorising is in part descriptive, but also that the main debate about vagueness, and the theories produced, *are* both descriptive in this way.

2iii. What of the prescriptive project?

What do we make of a 'prescriptive' project? As mentioned, we can take this to be a theory that tries to answer the 'how should we make distinctions with vague language?' question. The primary sort of prescriptive theory I have had in mind is one whereby the goal of philosophical inquiry is not to model features of the vague language we actually use, but rather give a theory that gives details of a better way to conduct our language; a paradox free logic and semantics, taking inspiration from Frege (1879, 1903), or Burgess and Plunkett's 'conceptual ethics' (2013). To this kind of theory experimental work, at least designed to enlighten us about competent language use, is of no use. But I don't think it's a very interesting project, nor what many philosophers are setting out to do in the discussion of vagueness.

There are lots of ways we could produce a prescriptive theory of this sort. Perhaps the most obvious is to follow Frege's path and precisify our language. We can avoid the sorites paradox, if we just arbitrarily place (sensible) boundaries in the extension of all vague language. We could get an 'ideal language', with a classical semantics and logic, and be free from the threat of paradox. Alternatively, to avoid the vagueness of colour predicates for example, we might say we should dispense of every colour predicate except 'red'. Everything is truly 'red', and we no longer have sorites problems with colour. We could do similar with other categories, height, wealth, age and so on. The resulting theory doesn't leave us with a very expressive language, but it's paradox free, and certainly simple and parsimonious. Or, we could keep our current lexicon, but suggest we treat them as if there is an unknown sharp cut-off to their extensions (regardless of their actual semantics and logic). This would be a prescriptive version of epistemicism, suggesting we treat our language as if it were precise, and maintain classical logic and semantics.

Any one of these theories might do the work desired of a prescriptive theory, providing us with a logic and semantics free of the problem of vagueness. But this kind of project gets off on the wrong foot. As conceded in the Burgess and Plunkett quote at the start of this section, paradoxes in natural language don't actually cause many problems; speakers communicate perfectly well, so why consider our concepts 'defective' at all? On the contrary, many seem to think that we couldn't operate with a language that was as precise as that developed by Frege. Williamson, for instance, says that 'if classical logic and semantics apply only to perfectly precise languages, then they apply to no language which we can speak' (1994: 2). The theories sketched above offer replacements for our concepts, re-definitions that leave them paradox free. But that isn't very interesting. There seem to be several ways we could achieve that aim, but what's the motivation to, given that speakers aren't driven to sorites paradoxes in everyday speech and thought. What *was* interesting in the first place is how speakers tend to stay paradox free with the vague language they operate with, and it is this that the philosopher should and, as I argued, *does* seek to explain.

I conclude then, that the goal of philosophical inquiry is, and should be, to accurately model the phenomenon of vagueness. Furthermore, I claim that if that is the goal, experimental data is relevant to theorising. These are the two premises of my argument, and from them I conclude that there is a relevant place for empirical work in the philosophical study of vagueness, at least in theory. In the following chapter, I will examine how people have tried to conduct empirical work, criticising some practical elements of them, and their theoretical foundation. Before moving onto that, I should defend the approach from some criticisms or worries that may arise at this stage.

3. Objections to the Experimental Approach

There are lots of good reasons why one might be suspicious of experimental philosophy, and perhaps especially when it comes to issues concerning logic. Hopefully the first section will have put some concerns to rest, showing that as we are concerned with modelling, empirical work can be relevant in theory. But there are still some issues that might need to be overcome before we can move on to thinking about how to go about conducting experiments. There is now a fair amount of literature criticising experimental philosophy. However, as discussed at the outset of this chapter, there is more than one sort of experimental philosophical project, and much of this literature isn't directed at the approach I'm defending here.

Most of the discussion is directed at the *negative* project, discussed above, which aims to undermine the use of intuitions as evidence in philosophical theories and arguments. Understandably many philosophers are resistant to this suggestion and seek to defend traditional philosophical methodology. For instance, Sosa (2007) objects to the methodology of studies which report cross-cultural differences in intuitions a means to argue against the use of intuition in philosophy (in particular Weinberg et al 2001, discussed above), arguing that subjects from different cultures are likely to understand the relevant thought experiments differently, on account of different background assumptions developed in their differing cultural and social upbringings. Ludwig (2007) seeks to argue that what subjects report in studies are not genuine intuitions, based on competent application of concepts, but some other kind of judgement, and that it is the task of the philosopher (from the armchair) to identify which judgements are intuitions which reflect competent concept application. Kauppinen (2007) argues that studies report unreflective judgements, and that philosophy uses more epistemically significant *reflective* intuitions. On a different tack, Cappelen (2014) argues that philosophers don't use intuitions in their theorising, and thus the experimental attack on them misses the mark altogether. There are many more in this vein, see Horvath and Grundmann's *Experimental Philosophy and its Critics* (2012) for extended discussion.

These sorts of objections are clearly not applicable to the project I have in mind here. The experimental approach to vagueness is not aimed at evaluating the use of intuition in philosophical theory, but rather informing that theorising. Of course, it is assumed that philosophers don't necessarily have a privileged access to understanding the workings of natural languages, so perhaps shouldn't rely solely on their intuitions in that respect. But, as argued, as philosophers are engaged in a descriptive project, they shouldn't assume they do have such privileged access (as we saw Keefe didn't, for instance). However, I will deal with an objection found in the literature that could reasonably be raised here.

3i. The expertise objection

This worry is the so-called 'expertise objection', which criticises the experimental approach for the use of non-experts in studies. This objection is also standardly discussed in relation to the negative project, as seen in Horvath for instance:

> As a matter of fact, the existing experimental studies cited in support of [the argument against the use of intuitions] were almost all conducted with subjects who had no or only very little philosophical expertise. Well-trained professional philosophers, however, can reasonably be expected to have intuitions about hypothetical cases that typically do not vary with irrelevant

> factors, at least not significantly and pervasively, as the [argument against the use of intuitions] requires. (2010: 464)

In this case, the push of the objection is that variation in intuitions of lay-people isn't problematic, as we can expect the intuitions of expert philosophers to be much more reliable, and those are the ones we use for theorising. Put like this, the objection isn't concerning for our purposes, as I'm not defending the negative project. But related concerns can put pressure on the approach to vagueness here.

The concern might be that lay-people aren't usually experts on logic (or the philosophy of language), and without training tend to be quite bad at it. So, someone one might fairly ask: why should we ask the folk about vagueness, when they're poorly equipped to tackle issues in the philosophy of logic and language? The way Knobe and Nichols see the objection might be closer:

> It's true that we are concerned with questions about common-sense concepts. The point is just that philosophers can use those very concepts—the ordinary common-sense concepts that people employ every day—with a precision and subtlety that ordinary people can't quite achieve. For the philosophers are specially trained to draw fine distinctions and to think carefully; and philosophers bring these skills to bear on uncovering the true nature of our common-sense intuitions. (2007: 9)

So, the idea is that normal people haven't had the right kind of training to give responses that are useful in experiments. If we want to know what the right model of vagueness is, then asking people who don't haven't have philosophical training isn't the right thing to do, only philosophers can think clearly enough about our concepts to provide intuitions relevant to theorising. Knobe and Nichol's own suggested response is that maybe we would see a difference in expert and non-expert intuition, but that that needn't be a problem; on the contrary they 'would love to know more about the ways in which philosophers differ from ordinary folks'. This is fine from their perspective on experimental philosophy as a psychological project. As discussed, they see the purpose of experiments as aiming reveal the workings of the mind, and why people have the intuitions they do. So, if it turns out the philosophers' expertise, and perhaps the lay-person's lack of theoretical contamination, affect their responses to thought experiments, then that is an interesting result for the experimental program.

Such a response isn't satisfactory for the experimental approach to vagueness, where we're trying to investigate the competent use of vague language; in this case, finding a significant difference between experts and lay-people would be problematic. If there were a reliable

difference in judgements of philosophers and lay-people, it might be because lay-people aren't well enough equipped to make judgements in the studies on vagueness (although, other explanations may well be possible). But there is a good response here. Ripley, in defence of experimental philosophical logic in general, gives the following reply:

> Untutored speakers cannot be assumed to be experts on the proper logical theory of vagueness; they can be assumed to be experts on whether they take anybody to be neither tall nor not tall (and if anybody, who?). It is precisely questions like the latter that are in play here. (2016: 531)

The point is that whilst lay-people aren't experts on the philosophy of language or logic, they are (or should be) experts on the competent application of the language being investigated. The experimental approach here suggests that we can learn about the right theory of vagueness by learning more about the language use of speakers, we are only investigating intuitions/opinions about the competent use of concepts, and accordingly the only training necessary is that of competent language usage.

This draws out an important distinction for experimental philosophy more generally, that between conceptual theorising and conceptual usage. The expertise objection may well have traction here if experiments aimed to record subjects *theorising* about concepts. For instance, a study might be directed at whether speakers find the that there *is* a sharp but unknown boundary between being 'rich' and 'not rich' counter-intuitive, as many philosophers do. The study might ask 'true or false: there is a number of pounds an Englishman can possess such that in possession of that amount they are rich, but a penny less and they are not rich'. This question invites subjects to theorise about the semantics of 'rich', the answer they give will be based on what they think the structure of a concept like 'rich' is following some reflection. But this isn't what we want to do, we want philosophers to do the reflecting and theorising. Our studies are meant to gather information on linguistic *usage*. That all speakers believe that there is or isn't a sharp boundary in the isn't good evidence for or against the epistemic theory. What would be good evidence is that the competent use of vague language does suggest there is one; for instance, one might think a study of the sort conducted by Bonini et al discussed earlier could provide such evidence.

The distinction between conceptual usage and theory can be seen in the psychological literature. For example, in Hampton and Passanisi (2016), it was found that what subjects considered to be the most important features of a category didn't map onto what subjects thought were the best exemplars of a category; the features a subject reported as the most

important for categories like 'insect' or 'sport' wasn't a good predictor what they reported were the most typical examples of insects or sports. One way we might think about what is happening here is that in listing what they thought were the most important features of some category, the subjects were *theorising* about concepts; they were reflecting on the important criteria for falling under a category. On the other hand, in deciding on whether an object is a good exemplar of a category, they are displaying competent application of a concept. The distinction has also already been made in a different area of the philosophical experimental literature. In criticism of the Machery et al studies on intuitions about reference discussed earlier, Marti argues that Machery's study invites subjects to theorise about reference, rather than test proper language use:

> [Machery et al.] test people's intuitions about theories of reference, not about the use of names. But what we think the correct theory of reference determination is, and how we use names to talk about things are two very different issues. (2009: 44)

Marti is arguing that we don't want to find out how people think reference works, we want to see how people use language to try figure out how reference works. In the same vein, we don't want to invite subjects to theorise about the semantic features of vague concepts, but rather want to look at patterns of competent usage, to see if might tell us something about the semantics. As Devitt and Porot write 'the right response to armchair philosophy is not to pull up more chairs for the folk' (2017: 2). This distinction shows why the expertise objection isn't problematic for the approach suggested here. It's true average speakers haven't been philosophically trained, but as experiments seek to test competent linguistic usage, the only expertise needed is in language use, which any fluent speaker will possess.

It's also interesting to point out some research carried out on the effect of expertise in experimental philosophy. Several studies seek to test if there is a difference in responses between experts and non-experts, and some also to see if that the difference is the expert responses are better (for the objection might require not only that there is a difference, but that experts give better responses). These have been conducted in various areas, but we can point to a few that investigate linguistic intuitions in particular. Culbertson and Gross (2009) and Sprouse et al (2013) test for an expertise effect in syntactic judgements, i.e. if some sentence is syntactically acceptable or not, and in both cases, it is found that expertise didn't make a significant difference. In a more direct study, Machery (2012) tested for the effect of expertise in studies investigating the reference of proper names, finding in one case that experts were more likely to have 'Kripkean' intuitions than non-experts (i.e. intuitions that

suggest a causal-historical theory of reference), and that experts are more likely to have descriptive intuitions in another, and takes this variation to undermine the expertise objection.

So, the expertise objection is not a big challenge for the project here. Fluent speakers have all the expertise required for a study on the competent use of vague language. And furthermore, studies designed to look for the difference between experts and non-experts in linguistic intuitions suggest that there no problematic difference.

4. Conclusion

We have now seen why the experimental approach can be useful for the philosophical investigation of vagueness. The question philosophers seek to answer is a descriptive one, aiming to model competent language use, and as it is a descriptive project, empirical data can in theory inform the search for the right answer. There is also no real threat to this approach from what seems a pressing worry, the expertise objection.

The biggest challenge for the project I'm defending is designing a study which can deliver genuinely significant results, and then interpreting those results in the correct way. Even if experiments can in theory be informative, how do we go about actually finding data which can help progress the debate? This is certainly difficult. In the following chapter, I will examine and criticise studies investigating vagueness. I'll argue many are flawed in their design and methodology, but also importantly in their theoretical foundations. For the experimental approach to vagueness to be successful we need a carefully worked out picture of what theories of vagueness would predict about competent language use, and studies up to date have made somewhat simplistic assumptions about this, assumptions I will correct later in the thesis.

Chapter 3 A Survey of Some Recent Experiments

Having established that there is the potential for an interesting role for experiments in theorising about vagueness, we might ask what, if anything, studies have taught us. In this chapter I will review three studies which use empirical methods to try shed some light of competent vague language use. I've selected three that demonstrate different experimental designs, to show the range in ways one might investigate vagueness. In each case I'll briefly overview the methodology, results and the interpretation offered, and in each case, I'll argue that the conclusions of the authors shouldn't be accepted. Whilst I do broadly endorse the experimental approach to vagueness, great care must be taken to ensure we get genuinely relevant results. There are a number of things required for a successful study, I'll outline a few themes now that we should keep in mind when evaluating the below.

Methodology and design: Is the design of the study appropriate for investigating vagueness and competent language use? A study should be carefully thought out, to ensure there is nothing that might skew or obscure the result

Statistical methods: Are the reported findings supported by appropriate statistical methods? Following the example of the social sciences, in order to satisfactorily conclude that a study has made a finding, we need to know more than merely that *x*% of participants report *y*. Statistical tests of significance measure how likely it is the observed data is the result of chance, and it is important to use them appropriately in order to conclude that there are genuine patterns in the data set

Cognitive predictions of theories: Do the authors have a good account of what sort of results different theories of vagueness would predict in the study? To be able to draw any conclusions from experimental data relevant to what kind of theory of vagueness is right, we need a well-developed theory what results are consistent with particular theories of vagueness. We cannot say whether some result supports supervaluationism for instance, unless we know what supervaluationism says about competent language use.

There are undoubtedly many more complexities when it comes to successful experimental conduct, but these are a few basic themes, and we'll see the studies to be considered here fail to cover all of these bases satisfactorily. Whilst I endorse the overall direction being taken in these studies, until the issues above are dealt with I don't think progress can be made.

It should be pointed out that there is of course more experimental literature on vagueness than I am dealing with here- it better to examine fewer interesting studies in depth to get a clearer picture of the problems that can arise, than to cover more less thoroughly[2].

1. Bonini et al

I begin with the study briefly discussed in the previous chapter, conducted by Bonini et al (with Osherson, Viale and Williamson) (1999). Bonini et al report the results of a series of experiments designed to investigate the mental representation of vague language. In each study the participant is asked to state the point at which there is a boundary in the extension of a vague predicate, such as the smallest height at which someone could be considered tall. One group were 'truth-judgers', and one group were 'falsity-judgers'. The truth judgers were asked to pick out the last point at which some predicate truly applies, and the falsity judgers were asked to pick out the last point at which some predicate falsely applies. Here's an example question from the first experiment:

When is it true to say that a man is "tall"? Of course, the adjective "tall" is true of very big men and false of very small men. We're interested in your view of the matter. Please indicate the smallest height that in your opinion makes it true to say that a man is "tall"

It is true to say that a man is "tall" if his height is greater than or equal to ___ centimetres

An analogous question was asked of the falsity-judgers. Over 6 of the studies reported by Bonini, similar questions were asked about other vague predicates, such as 'long' and 'old'. The idea behind the study is that we can learn about the way vague language is mentally represented by the differences between the answers given by the truth-judgers and falsity-judgers. For instance, we might think that if vague language is 'gappy', i.e. there is a truth value gap between negative and positive ranges of application for a vague word, then the average response of the truth-judgers will be significantly higher than the average response of the falsity-judgers; that the average value given for the shortest a man can be whilst being *truly* tall will be significantly higher than the average value given for the tallest a man can be whilst it is *false* he is tall, creating a gap between the two.

Bonini et al give interpretations of what some theories of vagueness would predict about the distance between the average response of truth judgers and falsity judgers. They suggest that a gap theory would predict the average value of truth-judgers to be significantly higher than

[2] Some interesting studies I couldn't cover in depth here include Ripley (2011), Sauerland (2011), Egré and Zehr (2016).

that of falsity judgers and that a glut theory (one that claims there is overlap between positive and negative ranges of application of vague language) would predict the average value of truth-judgers to be significantly lower than that of falsity-judgers. For a fuzzy theory, they suggest it depends what degree of truth is taken to be sufficient for assertion/denial. If we take *d* to be that value, and sentence P is assertible where it is true to at least degree *d* and deniable where it is false to at least degree *d* (i.e. true to at most degree 1 – *d*), then if *d* is greater than 0.5 (which is presumably should be) we can expect to find a gap between the range of assertion and range of denial, i.e. the lowest height at which it is assertible that someone is tall will be higher than the highest height it is deniable that someone is tall. Thus, Bonini et al take it that the average value of the truth-judgers will be significantly higher than that of the falsity-judgers.

The theory that Bonini et al seek to defend is epistemicism, and they formulate the following hypothesis:

Vagueness as Ignorance: S mentally represents vague predicates in the same way as other predicates with sharp true/false boundaries of whose location S is uncertain

In order to test this hypothesis, Bonini et al conduct a seventh study, which uses similar questions as those for vague predicates for precise ones which have sharp boundaries that the subjects are unlikely to know the location of. For instance, the following:

A man is not at least average age for an adult Italian if his age is greater than or equal to ___ years

A man is not as old as average for an adult Italian if his age is less than or equal to ___ years

These are meant to be analogous to the questions for truth-judgers and falsity-judgers respectively. The authors intend this to be a method of how vague predicates are mentally represented in comparison to precise ones which have unknown (yet knowable) sharp boundaries. As we'll see, it is argued that that for other predicates with uncertain true/false boundaries, the average value of the truth-judgers will be significantly higher than that of the falsity judgers (and produce empirical evidence in support of this), and thus the same will be true of vague predicates if epistemicism is true.

1i. Results

Bonini et al conducted six studies using twelve different vague predicates, generating 33 different trials in total. They report that gaps were found in every case, with only two not reaching statistical significance following a Mann-Whitney U test (being mountain in study four and study six). Table 1 below shows average values for truth and falsity judgers for 'x is tall' across four of Bonini et al's studies.

'x is tall'	Trial 1	Trial 2	Trial 3	Trial 4
Truth Judgers	178.30cm	179.55cm	181.49cm	170.28cm
Falsity Judgers	167.22cm	164.13cm	160.48cm	163.40cm

Table 1: Average values given by truth-judgers and falsity-judgers for 'X is tall' in Bonini et al

In a seventh study, Bonini et al test precise predicates with unknown but knowable sharp boundaries, including 'average age for an adult Italian' and 'average height for a 30-year-old Italian'. Five of the six tested showed statistically significant gaps (only the grams of alcohol per litre of blood required to be considered inebriated according to the highway code wasn't significant) again following a Mann-Whitney U test.

Bonini et al take the evidence from the first six studies to discredit glut theories of vagueness, and to be consistent with the gap theories, and fuzzy logic. They take study seven to show that the mental representation of precise predicates with unknown sharp boundaries will generate gaps between 'truth-judgers' and 'falsity-judgers', and thus that the gaps found in the first 6 studies are also consistent with epistemicism.

There is no obvious reason precise predicates should generate gaps, for either the 'vague' or not 'vague' predicates, but Bonini et al suggest that it could be explained by a preference for errors *omission* than of *commission* (citing some psychological literature is support of this, Ritov & Baron 1990, Spranca et al 1991). An error of omission would be to make a mistake by *withholding* a predicate when it truly applies, i.e. miss out part of the range of heights to which 'tall' truly applies, by reporting that the shortest you can be whilst being tall is a greater height than it actually is. An error of commission would be to make a mistake by *applying* a predicate when it is false, i.e. to include more than the range to of heights to which 'tall' truly applies, by reporting that the shortest you can be whilst being tall is a lesser height than it actually is.

If there was a bias towards an error of omission rather than commission, then speakers may be reluctant to give a value close to where the unknown sharp boundaries lie in the extension

of vague predicates (for fear of falsely committing to the predicate), driving the judgers towards clear truth and clear falsity away from the boundary, creating an apparent gap between the two. The idea is that when asked for the shortest someone can be whilst being truly tall, a speaker will be inclined to give a height where they are more confident that 'tall' applies, rather than get closer to the sharp boundary and risk claiming you can be tall at a height at which it is in fact false. As Bonini et al claim: '…this would induce reluctance by truth-judgers to descend far down the height-continuum, and reluctance by falsity-judgers to ascend too high. Gaps result.' (1999: 9).

So Bonini et al take the raw empirical data to possible support a gap theory, degree theory or epistemic theory of vagueness, whilst discrediting glut theories. They then give purely theoretical arguments against gap theories and fuzzy theories (based on arguments in Williamson 1994) to argue they are unsatisfactory, and thus conclude that the studies support epistemicism: 'This supports the conjecture that the psychological interpretation of vagueness rests on the assumption of a sharp but unknown boundary' (1999: 12).

1ii. Criticising Bonini et al

It is clear that the conclusion of Bonini et al cannot be accepted, as noted in various places e.g. Hampton (2007), Alxatib and Pelletier (2012). There are two methodological issues highlighted by Serchuk et al (2011) which conclusively undermine the study, based on the questions used and the statistical methods. But even beyond this, I don't this design is the right sort to bring us meaningful results, the questions presuppose the existence of sharp boundaries.

Ambiguous Questions

Serchuk et al point out that there is a potential problem with the questions used in Bonini et al's study. Consider:

Please indicate the smallest height that in your opinion makes it true to say that a man is 'tall'.

It is true to say that a man is 'tall' if his height is greater than or equal to ___ centimetres.

The first sentence asks to indicate the *smallest height* that makes it true for a man to be tall, this is asking for a necessary condition for tallness. However, the second sentence only asks for a height that is sufficient for tallness. For example, it is true that a man is tall if his height

is greater than or equal to 2 meters, however this is certainly not the smallest height that makes it true for someone to be tall (if there is such a height).

Serchuk et al were concerned that this shift from asking for necessary conditions to just sufficient conditions could skew the results of the experiment; if subjects were only actually asked to give a height sufficient for tallness or for not-tallness, then truth-judgers may give greater heights, and falsity-judgers may be lower heights, making gaps more likely. Serchuk et al test this hypothesis. They attempt to replicate the study in Bonini et al using their questions, and then also conduct a revised study with improved questions. In the revised study they ask only for necessary conditions, for instance:

What is the smallest height a man can be so that he is still tall enough for it to be

true to say that he is 'tall'? ___ feet and ___ inches

The replication and the revised study used the vague predications 'tall', 'old' and 'long'. Again, in both cases there were 'truth-judgers' and 'falsity-judgers'. The replication study found gaps between the average values of the sets of judges, the same result found in Bonini et al. However, in the revised study, the gaps disappeared. There was no significant difference between the average values for 'old' and 'long', but interestingly a statistically significant glut was found for 'tall'. Serchuk et al take this to undermine the result in Bonini et al:

> In our view, our data only license the conclusion that the ambiguity created the gaps observed by Bonini et al.; we do not believe that data from this kind of experiment, revised or otherwise, can reliably be used to argue for any theory of vagueness. (2011: 556)

It certainly seems decisive in at least as far as the particular result Bonini et al found; the question asked was not precise enough , and Serchuk et al have shown that the gap effect disappears when the question asked is revised.

Statistical Methods

Another methodological issue discussed by Serchuk et al is the statistics used by Bonini et al. As noted above, Bonini et al report that they found statistically significant gaps between the average values of truth-judgers and falsity-judgers in almost all cases, following a Mann-Whitney U test. As Serchuk et al (2011: 557), as well as Bonini et al, this test compares the *medians* of two sets of data. However, Bonini et al frame their hypothesis in terms of different *means*. Consider the following passages:

> On the gap hypothesis x is appreciably larger than y so we expect to find that x >> y (x, y are the average estimates of [truth-judgers] and [falsity-judgers] , and x >> y means that x is more than slightly greater than y). (1999: 3)

> The average responses to these queries are shown in Table 1, and reveal substantial gaps between the range of values in which the target sentence is deemed true, and those for which it is deemed false. For example, the range of indeterminacy for "old man" is ten years, which is more than 15% the size of the region in which the predicate is judged to apply falsely. Gaps of similar size show up throughout our studies. All three gaps are statistically significant according to the Mann-Whitney U-test (a nonparametric statistic suggested by the unequal variances that appear for some of the items in our studies). (pp. 5)

Bonini et al cash out the predictions of the theories of vagueness in terms of the average values of the judges, and the averages shown their tables are the *means* of truth-judgers and falsity-judgers, which in the same tables they suggest are statistically significantly different. But as the statistical test they used compares medians and not means, we do not know if there is significant difference between the means shown, we only know that there is a statistical difference between the medians in the data set. As the hypotheses laid out by the authors are based on the average values of the respective judges in the studies, the statistics used cannot be in support of those hypotheses.

As they note in the above quote, Bonini et al selected a Mann-Whitney U test on account of the unequal variances in the data. They were unable to use a *parametric* statistical test, such as a *t-test* to compare the means, as this requires that the data falls in a normal distribution i.e. it makes assumptions about the variance and standard deviation. The data found by Bonini et al did not fall in a normal distribution; this means that the responses of subjects falls in a wide range around the mean. Serchuk et al note that the data they found when replicating the study had a large standard deviation, so they also couldn't use parametric tests, and suspect that this is the same reason Bonini et al didn't (2011: 558).

If we want to learn something about vagueness from empirical work, it is essential that appropriate statistical methods are used. We cannot learn anything from simply eyeballing raw data, statistics are how we can detect trends within that data. Bonini et al have drawn out hypotheses about the average estimates in their studies, and then claimed to have supported those hypotheses. But their statistics only tell us the difference between the median values within their two groups, not about the difference in central tendencies, which is what their hypotheses required.

Theoretical Concerns

I take the two methodological problems with the study to decisively show that we cannot accept any conclusions from the data in Bonini et al; their conclusions relied on inappropriate

statistics, and the gaps observed were likely the result of poorly written questions. I also think that the general design of the study isn't the right sort to show us something about vagueness.

Serchuk et al also accuse the study of being question begging, as their questions assume the existence of a sharp boundary in the extensions of vague predicates, which they claim is in fact 'the most worrisome problem' (2011: 549). Consider: 'Please indicate the smallest height that in your opinion makes it true to say that a man is 'tall''. This question presupposes that there is some such height, or that it would be reasonable for participants to claim what they think that height is. Serchuk et al take this to be a very serious issue with the design:

> This makes the design question-begging. It assumes that epistemicism is true, because if epistemicism is not true then the question is nonsense; moreover, even if epistemicism is true, there is simply no way for a participant who does not believe that vague predicates have precise boundaries to respond. (pp. 550)

I agree that this is problematic, but I don't think it's quite right to say that the study is question-begging. That the questions asked presuppose sharp boundaries doesn't mean that it was any more likely that Bonini et al would find evidence for epistemicism; the authors do not take themselves to be discovering anything about the location of actual metaphysical sharp boundaries, nor demonstrating there are any. The result they claim to find, that there are statistically significant gaps, is just as consistent with their hypotheses for gap theories and fuzzy theories as it is for epistemicism.

So, the problem isn't that Bonini et al have *begged the question* here. But I do agree that the right way to investigate vagueness is not to ask speakers where the sharp boundaries lay, at least without some qualification. A central feature of vagueness is that it seems there are no such boundaries, acknowledged even by theories that suggest boundaries do exist, so it's unclear whether we can expect to learn something about competent language use by asking speakers where they are. This issue is amplified by the fact that participants didn't have the option to give a response like 'I don't know' or 'there isn't such a height'. Rather, participants were forced to give a height that was the 'shortest someone could be truly tall in their opinion'. With this kind of questioning it's not clear we can learn anything about competent vague language use, and participants may well give chaotic or unreliable responses; the large range of answers reported by both Bonini et al and Serchuk et al may well be the result of participants being forced to try to identify a value they do not even believe exists. Given this problem, we cannot very reasonably take any predictions from theories of vagueness about the results of a study like Bonini et al's. For instance, a 'gappy'

theory of vagueness needn't make any prediction about how the average estimates of truth-judgers and falsity-judgers will compare when they are forced to report where they think the sharp boundaries are, as gap theories take it there are no such boundaries.

Another concern is whether theories of vagueness can be taken to make a prediction about the comparison of two separated groups, in the way they're compared in Bonini et al. What is important to theories of vagueness is how *individuals* mentally represent vague concepts. However, comparing the values of truth-judgers and falsity-judgers needn't tell us anything about how those individuals represent the vague concepts they are judging. When a truth-judger reports that height *x* is the smallest someone can be whilst it being true that they are tall, we don't get a full picture of how they as an individual mentally represent 'tall'; if they would give a value greater than or less than *x* as a falsity-judger for example.

The point is more poignant given the way the 'gaps' disappear when the questions aren't ambiguous, and the large standard deviations in the data. Finding a reliable significant difference between the means of truth-judgers and falsity-judgers could at least give us indirect evidence of how individuals represent vague concepts; if there were a reliable pattern found we could arguably generalise from the groups to individuals. But instead we find that subjects are likely to give responses over a wide range, without any significant difference to that value if they are truth-judger or falsity-judger, and we learn nothing of how any of those subjects individually represent vague concepts.

The study by Bonini et al then, doesn't teach us anything about vagueness. It is a good example of the difficulties facing the experimental approach; it's important to take great care over the design, to ensure that the methodology is appropriate practically and theoretically.

2. Alxatib and Pelletier

Alxatib and Pelletier (2012) report the results of a study initially motivated by issues with Bonini et al. They are critical of Bonini et al's rejection of gap theories of vagueness, and set out to generate data relevant to an issue with some of the theoretical arguments against gap theories in Bonini et al. We're not interested in their concerns with Bonini et al here, but Alxatib and Pelletier go on to report some other interesting results, which they then use to support a novel pluralist framework.

Alxatib and Pelletier presented participants with a synthesised image of five men standing in a police line-up. The wall behind gives people a means of measuring their heights, and it is supposed to appear that they stand at 5'4", 5'11", 6'6", 5'7", and 6'2". Each man is given a number, 1-5. Participants were given twenty true/false questions in a random order, describing the men in the picture, such as the following:

#1 is tall

#1 is not tall

#1 is both tall and not tall

#1 is neither tall nor not tall

Analogous questions were asked of numbers 2-5. For each statement, the participants were asked to respond with 'true', 'false' or 'can't tell'.

Man #2 in the line-up, at 5'11", is meant to represent the borderline case of 'tall', being the median height. Alxatib and Pelletier are most interested in the way this man is described, suggesting it can be significant for gap theories or glut theories of vagueness, and they have supervaluationism and subvaluationism in mind in particular here (pp. 299). Whilst they don't mention any particular hypotheses at the outset, they do suggest that the truth of the *both* question (i.e. #2 is both tall and not tall) suggests a truth value glut, and the truth of the *neither* question (i.e. #2 is neither tall nor not tall) suggests a truth value gap. They take it subvaluationism predicts assent to *both*, and supervaluationism predicts assent to *neither,* and find assent to both, Alxatib and Pelletier recommend a framework which makes use of both supertruth and sub-truth.

2i. Results

Alxatib and Pelletier report a number of interesting (looking) results. They find some assent to *neither* and *both* for man #2. 44.7% of participants reported that '#2 is both tall and not tall' was true. 40.8% reported it was false, but this is lower than the percentage that thought the contradiction was false for the other men. The suggestion is that this lends some support to a glut view, in which borderline cases are contradictions. 53.9% of participants reported that '#2 is neither tall nor not tall' was true. 42.1% reported that it was false, which again is lower than percentage response of 'false' to the *neither* question for the other man in the line-up. Alxatib and Pelletier also suggest that true responses to *both* and *neither* correlate with each other, with 53.7% that reported *neither* was true also reporting that *both* was true, and

64.7% that reported *both* was true also reporting that *neither* was true. This result is significant for their theory:

> What we want to highlight is that *neither*, whose truth can justify a truth-value gap, coincides in many cases (more than half!) of borderline-height with *both*, which, when true, suggests a truth-value glut. (pp. 314)

But it's worth noting at this point that Alxatib and Pelletier don't report that this is a statistically significant correlation, but only give us some percentages to eyeball, a topic I'll return to.

The authors also report interesting results when it comes to responses to *both* and *neither* considering the judgements on the individual conjuncts/disjuncts that make up those statements. 55.9% of participants that thought #2 was both tall and not tall, answered that one of the individual conjuncts was false and that the other was true, 32.4% thought that they were both false and only 2.9% answered true to each conjunct individually. 51.2% of the participants that reported that #2 was neither tall nor not tall, thought that one disjunct was false and the other true while only 24.4% thought both disjuncts were false individually. Table 2 sums up the responses to the simple sentences given the judgment on the complex ones.

Response to *Both/Neither*	Judgements on conjuncts	% within answer type of *both/Neither*
True	Both True	2.9% / 2.4%
	Both False	32.4% / 24.4%
	Both Can't Tell	5.9% / 9.8%
	One True, One False	55.99% / 51.2%
	Other	2.9% / 9.8%
False	Both True	3.2% / 6.2%
	Both False	12.9% / 15.6%
	Both Can't Tell	0.0% / 0.0%
	One True, One False	74.2% / 78.1%
	Other	9.7% / 0.0%

Table 2: Percentage of response types given response to 'both'/'neither' in Alxatib and Pelletier

Here I haven't included the responses when the participants answered 'can't tell' to *neither* or *both.* This is because can't tell was used very little by the participants in response to those statements, and throughout the study in general. Only 9.2% of participants answered 'can't tell' for '#2 is tall', 7.9% for '#2 is not tall', 14.5% for '#2 is tall and not tall', and 4% for '#2 is neither tall nor not tall'. This is somewhat surprising, as borderline cases are standardly described as those which are very difficult to make a judgement about, so one might have thought 'can't tell' would be a popular response when judging a borderline tall man. Alxatib and Pelletier note that there might have been a bias against using the 'can't tell' option, participants might have thought it represented a 'regrettable lacuna' in their knowledge, meaning something like 'I give up' (pp. 309).

Another interesting feature we see coming out of this is that the majority of participants are giving classical assignments to '#2 is tall' and '#2 is not tall', i.e. judging one is true and the other false; overall, 63% of participants responded in this way. Whilst this classical assignment is the most common response, Alxatib and Pelletier take particular interest in a different response type; one where participants reported that *both* was true, but judged each individual conjunct as false. As mentioned this response was given by 32.4% of participants who judged *both* to be true, and it plays a central role in the framework the authors defend.

Alxatib and Pelletier suggest the results support a new theory of vagueness, which is pluralist, and makes use of pragmatic maxims. They suggest that vague predicates are ambiguous between two possible semantic interpretations; a super-interpretation and a sub-interpretation. Consider 'x is tall'. A super-interpretation is like a supervaluationists interpretation; 'x is tall' is true iff true on all admissible precisifications of 'tall'. A sub-interpretation is something like a subvaluationist interpretation; 'x is tall' is true iff true on one admissible precisification of 'tall' (and they make clear that they do have the familiar notions of supertruth and sub-truth in mind here, making use of *admissible precisifications*, e.g. pp. 320). The suggestion is that vague predicates can be interpreted either way, and how they are interpreted is governed by pragmatic maxims, in particular Gricean conversational maxims. For example, Alxatib and Pelletier suggest that the maxim of quantity requires that the strongest meaning is always taken, so for a straightforward claim like 'x is tall', this maxim suggests the *super-interpretation.* For a claim like 'x is tall and not tall', the super-interpretation would necessarily yield falsehood, but can be true on the sub-interpretation.

Alxatib and Pelletier use this framework to explain why participants would assent to the contradiction *both* whilst rejecting both disjuncts. The idea is that '#2 is tall' and '#2 is not tall' will both be super-interpreted, and as #2 is a borderline case, each is false. There are four different ways that the conjunction of the two could be interpreted: as super-tall and the negation of super-tall, as sub-tall and the negation of sub-tall, as super-tall and the negation of sub-tall, as sub-tall and the negation of super-tall (Alxatib and Pelletier take negation to be a function which returns a sentence's complement. The negation of super-tall returns borderline (i.e. sub-tall) or super-not-tall, the negation of sub-tall returns super-not-tall). Alxatib and Pelletier suggest that it is reasonable to use the Gricean maxim of quality, and thus won't choose an interpretation that yields an empty set. Thus, the only acceptable interpretation of the contradiction is last one, sub-tall and the negation of super-tall. This yields that the individuals that are sub-tall and either borderline or super-not-tall, which can only consistently be the borderline cases. On this theory it is consistent for participants to say each conjunct of the contradiction is false on its own, but a conjunction of them is true.

This is the author's preferred interpretation of the data. They argue supervaluationism couldn't adequately explain the results. They suggest the ideal *supervaluationist* response would be to assent to '#2 is neither tall nor not tall', whilst answering 'can't tell' to each disjunct individually, which is rarely seen. They allow that perhaps the supervaluationist could argue that all that really matters is assent to *neither*, but that the theory couldn't accommodate other parts of the data:

> …our supervaluation theorist may say, we should allow any answer to the disjuncts. The thing of importance is that they thought that '#2 is neither tall nor not tall' is true. Perhaps. However, it might also be noted that of the 41 subjects that answered 'true' to neither, 22 of them (53.7%) also answered 'true' to '#2 is both tall and not tall', which should not happen, according to supervaluationism. And 6 (14.6%) of them answered 'can't tell' to this, which also shouldn't happen according to supervaluationism. Thus only 13 of the subjects that answered 'true' to neither (31.7%) are acting as supervaluationists, even with this generous understanding of the way they might answer the atomic questions. (And recall that this means that only 13 of 76 subjects in the entire pool, 17.1%, were answering in accordance with supervaluationism.) (pp. 322-3)

The authors also argue that a degree theory couldn't explain the results seen here, even though it could account for the levels of assent to contradictions. On a standard degree semantics, the truth value of a conjunction equals the truth value of the conjunct with the lowest truth value. Thus, where P is true to degree 0.5 and ¬P is true to degree 0.5, the truth value of 'P & ¬P' will also be 0.5. As the contradiction '#2 is tall and not tall' doesn't come out entirely false, a degree the authors suggest that a degree theory could claim to explain the

levels of agreement with it. However, Alxatib and Pelletier reject a degree theoretic interpretation as they claim it cannot explain the wider patter of results:

> a closer look at how the same judgers, taken individually, responded to other queries reveals a recurrent pattern that the fuzzy approach cannot predict, namely, the pattern in which a borderline proposition, and its negation, are judged false, but in which their conjunction is simultaneously judged true. (pp. 321)

So, the authors take it that their pluralist/pragmatic theory best explains the data, but there are good reasons to resist this conclusion.

2ii. Criticising Alxatib and Pelletier

Whilst this study does bring out some results that *look* interesting, unfortunately I don't think we can say it teaches us much about vagueness. We can object to Alxatib and Pelletier on both theoretical and methodological grounds.

Use of Statistics (or lack thereof)

A striking feature of this study is that throughout most of the important discussion, statistics are very rarely used or mentioned. Alxatib and Pelletier aim to support a new theory of vagueness based on some certain results they see, but they cannot claim these response types are statistically likely. A statistical test of significance is designed to calculate how likely it is you'd get the particular results of a study if the *null hypothesis* were true; the null hypothesis claims that there will be no difference between two experimental groups, or a particular type of response will be no more common than any other. If such a test doesn't find significance, then you cannot rule out the possibility that the results were down to chance. We cannot really be said learn something from knowing x% of people gave response y. Eyeballing percentages isn't a reliable method of evaluating trends seen in the data set. If we want to seriously investigate vague language use, we need to much more rigorous. Statistical analysis gives us a way of evaluating the importance of x% giving response y%, based on other features of the data.

For example, take Alxatib and Pelletier's highlighted 'finding', that participants are likely to assent to a contradictory description of a borderline cases but dissent from each conjunct. A statistical test of significance would calculate how likely it is you'd get the number of participants responding in that way if the *null hypothesis* were true. Alxatib and Pelletier don't spell out any hypotheses explicitly, but in this case, it would be something like

'participants are no more likely to give that response type than any other'. As no test is carried out, we cannot conclude that the result seen is not simply due to chance.

This can be said of almost any interesting 'finding' from the study, as statistics are only mentioned in one case; in an earlier discussion on Bonini et al, where they report there was a statistically significant preference for denying a proposition than asserting its negation (pp. 308, *fn* 14). The study has been citied on several occasions as an example of finding a willingness to assent to contradictions for borderline cases. Whilst it's true we can say that 44.7% of participants reported '#2 is tall and not tall' is true, we cannot claim that a robust finding of study is that participants are likely to assent to contradictions; this would require the support of a statistical test, and given that the majority of participants didn't report it was true, I suspect it's unlikely to reach significance.

I think the same is likely to be true of other 'findings' of the study, and perhaps this is why the authors don't mention statistics. It is noted that 32.4% of participants that reported *true* for both reported each conjunct was false individually, which means 11 participants in total gave this response type. Based on these figures it seems unlikely that either people that response overall or people denying each conjunct whilst assenting to *both* would reach statistical significance. This certainly makes the central role this result plays in the study look very strange, but perhaps explains why statistical tests aren't mentioned[3].

Problems for their Interpretation

As discussed, Alxatib and Pelletier suggest as an explanation of their data a theory which makes use of both supertruth and sub-truth, invoking pragmatic maxims to account for which interpretation speakers use. A key reason for objecting to this interpretation has now been highlighted. The main motivation for this theory was to explain why participants would assent to *both* whilst dissenting from each conjunct. But as stated above, this result is not statistically significant, and does not warrant explaining; we cannot rule out the possibility the result was due to chance.

But even with statistics aside, one might ask why Alxatib and Pelletier thought it necessary to explain a response pattern seen given by 11 out of the 76 participants in the study, especially given that their theory cannot explain more common response types. The most common

[3] Interestingly, in an earlier version of the paper (published in the volume *Vagueness in Communication* (2011)), the authors do mention a single statistical test in support of the correlation between participants assenting to *both* and *neither* (2011: 30, *fn*13). For some reason this is omitted in the later publication.

response for those who assented to the contradiction was to assent to one of the conjuncts and dissent from the other. The theory offered by Alxatib and Pelletier cannot straightforwardly explain this result (as they claim 'x is tall' and 'x is not tall' should both come out false when super-interpreted), yet they offer no reason why they don't address this, nor justify why it is the less common response that takes a central role in the paper. It is unclear why we should be interested in a theory that is intended to explain a pattern that is uncommon and not statistically significant.

Beyond this, one might take issue with the adequacy of the explanation Alxatib and Pelletier do offer of their favoured finding. They claim to be invoking a model which makes use of both supertruth and sub-truth in the standard sense, that is truth on all and some admissible precisifications, respectively. But it's not obvious how true their model is to these familiar notions. For instance, standardly supervaluationism and subvaluationism evaluate *whole sentences* on their admissible precisifications. Alxatib and Pelletier do seem to gesture that they have this in mind at some points, e.g.:

> A hearer is to find the more suitable interpretation of the predicate from these two possible meanings, or just to say that there is no way to choose and the *sentence* is simply ambiguous. (2011: 319) (emphasis added)

If this were the case, then whole complex sentences would have to be evaluated on admissible precisifications, which is why both supervaluationism and subvaluationism preserve, e.g., the law of excluded middle, in the face of vagueness. But if this were the case then '#2 is tall and not tall' would come out false whether super- or sub-interpreted, as it is false on every admissible precisification. Alxatib and Pelletier don't take the semantics of complex sentences to work in this way though, and allow that both instances of 'tall' in the conjunction could either be super- or sub-interpreted. Perhaps there is some justification for that, as the model has a plural notion of truth, but it should be acknowledged that this is a serious deviation from the semantics of connectives in supervaluationism and subvaluationism.

One might also wonder why 'x is tall' should be judged *false* for a borderline case when super-interpreted, if the theory makes use of supertruth. When x is a borderline case of tallness, 'x is tall' isn't true or false, so it's not clear why it fits with their theory that participants do judge it as false. They best they offer is that 'x is (not) tall' 'is likely to be disagreed with, since [x] does not qualify as super-tall, or super-not-tall' (pp. 318). It isn't

clear why it's not the case that participants should answer 'can't tell' for each conjunct according to their theory, as they suggest *is* the case for supervaluationism later on (pp. 321).

We don't need to discuss the theory at length, as it is at best an *ad hoc* response to a result that is far from empirically robust, but it should be noted that the theory doesn't necessarily explain the one result of the study it was meant to.

The Predictions of Theories

Throughout the study Alxatib and Pelletier make suggestions here and there about what sort of response patterns would be support one theory of vagueness or another, or why some theories cannot explain the result seen in the experiment. But they lack a well-developed theory of what a particular theory of vagueness would predict from an experiment like this, and it can be seen that the suggestions they make are too simplistic because of this.

For instance, they initially take themselves to be comparing truth-gaps, and truth-gluts, and think that *neither* and *both* capture these:

> …*neither*, whose truth can justify a truth-value gap, coincides in many cases (more than half!) of borderline-height with *both*, which, when true, suggests a truth-value glut. (pp. 314)

The authors make they make clear from the outset the main 'gappy' and 'glutty' theories they are interested in are supervaluationism and subvaluationism (pp. 289-90), they don't discuss any alternatives. However, it cannot straightforwardly be taken that '#2 is neither tall nor not tall' should be judged true according to supervaluationism; as this theory validates all classical tautologies the law of excluded middle is always true, and *neither* would be necessarily false. Similarly, it cannot be straightforwardly taken that '#2 is both tall and not tall' should be judged true according to subvaluationism; this theory also validates all classical tautologies, and thus *both* comes out necessarily false.

Whilst it is true that supervaluationism and subvaluationism have truth value gaps and gluts of sorts, they are not characterised by the failure of the law of excluded middle, or law of non-contradiction. Bivalence fails for supervaluationism, meaning it is not that the case that every proposition is either true or false, however 'P v ¬P' is always supertrue. Subvaluationism allows that 'P' and '¬P' can simultaneously be true, however the conjunction 'P & ¬P' is necessarily false. There are stronger paracomplete and paraconsistent options available of course, according to with 'P v ¬P' can fail to be true, as can the law of non-contradiction, which are more obviously relatable to the suggestion that the truth of

neither and *both* indicate respective truth-gaps and truth-gluts. However, to suggest that supervaluationism and subvaluationism can be related to this suggestion clearly takes more qualification, but Alxatib and Pelletier offer none.

The same thought applied to their later suggestion of what a supervaluationist would expect from their data:

> …of our 76 subjects, only 4 answered 'can't tell' to the individual disjuncts and also answered 'true' to the neither question. So, this is not the correct way to look at the subjects' internal view of their choices. But in turn, this seems to show that people do not tend toward a supervaluation-style theory, since that is the theory that is naturally associated with this view. (pp. 322)

Again, we cannot straightforwardly suggest that supervaluationism predicts that subject would answer 'true' to *neither*, given that it is necessarily false. We need a carefully worked out theory of what theories of vagueness really would predict in a study like this if we were to make progress.

Alxatib and Pelletier's rejection of degree theories has a similar issue. As the authors do point out, there are some initial results that a degree logician could point to in support of a degree theory of vagueness. In particular, the truth value of a contradiction can reach up degree 0.5, we can explain the level of assent seen in the study to *both*, as well as why it peaks for man #2, the borderline case, and slopes down as the men get taller and shorter. Alxatib and Pelletier reject the theory as it cannot explain their favoured result, whereby participants accept both whilst rejecting the conjuncts:

> We think that proponents of this view argue in favour of the fuzzy approach without taking notice of how believers of contradictions—the truth-judgers of 'tall and not tall'—judge the truth of other related statements like 'x is tall' and 'x is not tall'. (pp. 321)

We can see now that this particular rejection is unjustified. Degree theories shouldn't be rejected on account of a response seen only by 11 of the 76 participants, and one which is not statistically likely. What's more, depending on how we cash out the predictions of a degree theory, we could get a picture that does sit well with the results seen. The authors offer no real theory of how to interpret the cognitive implications of a degree approach to vagueness, but if do flesh it out a little we can see that they are too hasty to reject such a theory.

For instance, if we take a model suggested by N. J. J. Smith (2010), where an agent's degree of belief in a proposition S is a function on the degree of truth of a proposition, and this degree of belief translates into the agent's *strength of tendency to act* as if 'S' were true. This

model will be examined in depth in chapter 6, but for now this brief suggestion gives us enough to illustrate how a degree theorist might explain the data here.

One way to understand 'strength of tendency to act as if true' in this case, is the likelihood to answer 'true' for any of the statements in the study, and if we do so we see that many of Alxatib and Pelletier's results are amenable to the degree approach. As already noted, the percentage of 'true' responses for *both* peaks at the median height, and decrease as the height increases or decreases. The median height, the borderline case, is where the truth value of *both* will peak according to the degree approach, so the strength of tendency to act as if *both* is true (in this case, the tendency to answer 'true') will peak at that point also. A degree theorist might also argue we see roughly the right levels of assent to the conjunction. 44.7% answered it was true, on the model we're currently considering, this would suggest its degree of truth was around 0.45, which is close to a paradigmatic borderline case. Similar stories can be told about the levels of assent to 'x is tall'. As the man in the line-up gets taller, the levels of assent to 'x is tall' increases, with 46.1% for the borderline case man #2. With the current model, one could suggest that this represents the truth value of each statement increasing, and the *strength of tendency* to act is each is true increasing with it.

With this model in mind, contrary to the suggestion of Alxatib and Pelletier, the degree theorist *can* explain the attitudes subjects take to the individual conjuncts as well as their attitudes to *both*. As already noted, we cannot reject the degree approach simply on the fact that 11 subjects reported each conjunct was false whilst assenting to *both*. Alxatib and Pelletier are correct this response pattern could be problematic for the degree approach, but only if it was a statistically likely response. As it stands, it is not even the majority of participants that assent to *both*, let alone the whole sample. What we see instead is a pattern that is amenable to a degree theory; that no particular response pattern is particularly likely. Where we are taking degree of truth to translate into strength of tendency to act, then if 'x is tall' and 'x is not tall' are borderline true propositions, there will be a middling strength of tendency to act as if each is true, which might come out as a mixture of responses to the individual conjuncts. This is the sort of pattern we do see in the results. The most common response for those assenting to *both* is to assent to one conjunct and dissent from the other. This is still compatible with the model we're considering, insofar as one combination isn't more likely than the other (and it isn't reported that one is more likely than the other).

I'm not suggesting a degree theoretic interpretation of the data here. Rather, I aim to draw out a lack of theoretical foundations in the study when it comes to interpreting the cognitive predictions of theories of vagueness. The author's interpretation of supervaluationism and subvaluationism is clearly underdeveloped, we cannot take assent to *neither* and *both* as evidence of either theory without some further qualification. The author's interpretation of degree theories is also underdeveloped, and with a more fleshed out account, a case could be made for a degree theoretic interpretation of their results. For the experimental approach to vagueness to succeed, it is necessary that we first do have a solid theoretical account of the cognitive frameworks that might come with a theory of vagueness, in order to be able to determine how well a set of data does sit with any theory. I will aim to take steps towards developing these kinds of frameworks in the following chapters of the thesis.

3. Egré et al

Egré et al (with De Gardelle and Ripley) (2013) report the results of a study which is focussed around the *hysteresis* effect. Hysteresis in general refers an object shifting from one category to another at some point along a relevant continuum, where the point of category shift varies depending on the direction travelled along that continuum. Such an effect is observed in many places. Raffman reports water as an example; ice will melt at 0 degrees, but water will freeze at -4 degrees (2014: 140). Here water is shifting category (frozen/unfrozen) along the continuum of heat, and the point of category shift changes depending the direction of travel along that continuum. In particular, hysteresis is intended to pick out the case where the category stated with persists longer in each direction. In the given example, water remains unfrozen past the point at which ice will melt, and similarly ice will remain frozen past the point at which water freezes.

Such an effect has been discussed in relation to the semantic categorisation of objects in a sorites series. Say we take a sorites series from red to orange. The suggestion is that if speakers are asked to start at one end of the series and categorise each shade in turn, they will shift category at some point, but where that point is will vary depending which end they start at (and thus which category they start applying first, red or orange). It might be that speakers will persist with the category they start with for longer in each direction, so that they switch from red to orange at a shade closer to *orange* than the shade at which they switch from orange to red; call this a *positive hysteresis.* Alternatively, it might be that speakers won't persist with the category they start with in each direction, so that the shade at which they

switch from red to orange is closer to *red* than the shade at which they switch from orange to red; call this a *negative hysteresis.*

Egré et al seek to investigate this effect, in particular in comparison to the findings of Raffman (2014) and Kalmus (1979). Kalmus conducted a colour categorisation task like that described, and found a negative hysteresis affect; participants were likely to shift colour category closer to the colour they were coming from. However, Kalmus and Egré et al suggest this might be the result of some 'forward-looking'. Participants were informed of the direction of the colour spectrum shown in advance, so they would have been anticipating having to change category, and this may have made participants more likely to shift earlier.

Diana Raffman (2014: chp. 5) reports the results of some experimental work, in support of a novel theory of vagueness she defends, which is a hysteresis-type study. In the normal condition, where participants where just shown a sorites series from blue to green (or green to blue) in order, she reports there was something like a negative hysteresis, as there was in Kalmus, but in this case the result cannot be explained by 'forward-looking'. In a slight variation, a *reversal condition*, Raffman reports a positive hysteresis. In the reversal condition, Raffman showed participants the sorites series from blue to green, and once they shift category (say to green from blue), she would switch the direction of travel along the series, showing them the shade previously seen (which they had just described as blue), and reported participants would then categorise it as *green*, and some that followed it.

In general, Raffman's study shouldn't be given significant attention. The study was conducted on only 19 people, and the discussion is very informal. Raffman only gives details of three participants, which are strong examples of the sort of result she's interested in, but there is far evidence of clear patterns. For instance, Raffman reports that in the reversal condition, two participants had a positive hysteresis of 8 shades, and one had a positive hysteresis of 15 shades, but the average was less than 3. Even if this average were 2.9, the remaining 16 participants would have averaged around 2.3 shades, which is not a very strong effect. We aren't given details of the whole study and patterns in the data (for instance, full details of why she concluded there is a negative hysteresis effect in the normal condition), Raffman only uses a couple of particular subjects that display patterns she's interested in.

Nevertheless, Egré et al intend to investigate whether a *positive* hysteresis effect is a robust finding following Raffman's reversal condition (despite both Raffman and Kalmus reporting a negative one in the general condition), and predict that they will find a positive hysteresis.

At the outset of the study, Egré et al suggest that a positive hysteresis would support an 'overlap view' of vagueness, that is a 'glutty' theory, 'since there is a range of shades that subjects equally call BLUE and GREEN depending on the context of the transition' (2013: 393).

Egré et al conduct two experiments in total. The first concentrates on hysteresis, and includes both a perceptual and linguistic task, based on a series of colour squares from blue to green, and yellow to orange. In the perceptual task, participants were shown three colours at once. Two of these were clear cases at the beginning and end of each series (i.e. yellow and orange), and the third was a shade from the series. The participant had to judge which of the two clear cases the third shade was most similar to. In the linguistic task, only one shade appeared at a time from the series, and the participant had to choose to describe it as either 'yellow' or 'orange' (or 'blue' or 'green' for the blue-green series).

In the second experiment, participants were again presented with a colour series one shade at time, but were also shown a sentence describing its colour, and had to either 'agree' or 'disagree' with the description. There were 48 different conditions in this test; either the yellow-orange continuum or the blue-green continuum could be used, the description could involve either of the two colours in the continuum, the description could be atomic or conjunctive, and the description could involve conjunction or not (i.e. 'the square is yellow' vs 'the square is not yellow', 'the square is yellow and orange' vs 'the square is yellow and not yellow'). Finally, the colours in the continuum could be presented blue to green (yellow to orange), in reverse, or in a random order. For each condition, participants were shown 15 colour squares, and were given each condition twice (meaning 1440 trails per participant!).

The aim of the second experiment was to investigate hysteresis; will the point at which a participant stops agreeing to 'the square is yellow' vary depending on the order of direction along the series. A secondary aim was to investigate the levels of assent to the conjunctive descriptions, predicting assent should be higher for the borderline cases in each series.

3i. Results

Egré et al report that in the first experiment, there was no significant order effect found (i.e. no negative or positive hysteresis) in the perceptual task, but an ANOVA (analysis of variance) showed a statistically significant *negative* hysteresis in the linguistic task, in line with Kalmus (and perhaps Raffman).

Egré et al consider three different possible explanations of this result. They suggest we could take the negative hysteresis to show that the vague concepts in the study *exclude* each other. One version of this view is an *epistemic* one; blue and green have a sharp unknown boundary between them. In this case, the authors claim a negative hysteresis could be explained by a preference for errors of omission over those of commission (2013: 402), similar to the notions we saw utilised in Bonini et al (1999). The idea would be that participants would be happily using (say) 'blue' at the start of the series, but as the colour shades become less blue and more green, the participant would be aware that the sharp boundary was approaching. The suggestion is that participants might prefer to shift to 'green' earlier sooner, and omit calling some blue shades 'blue', rather than persist and risk distorting their thus far correct application of 'blue', by falsely calling a green shade 'blue'.

A second version of the exclusion view offered is a supervaluationist gap theory; there is a range in between the truly blue and green shades of shades which are neither. In this case, the suggestion is that participants would start by accurately applying 'blue' to the clearly blue cases, but upon reaching the borderline cases, may feel that it is preferable to shift to a new category rather than distort their previous use of 'blue'; it may feel improper to suggest the shade is still blue once subjects reach the borderline cases, and seem useful to switch to 'green' to indicate the semantic status of the shades has changed.

Egré et al suggest an alternative framework which they prefer to explain the result, which uses the semantics expounded in Cobreros et al (2012). This framework is pluralistic, each vague word has a *strict* and a *tolerant* extension. Strict extensions underlap, and so have truth value gaps, and tolerant extensions overlap, and have truth value gluts. As we saw in the theory suggested by Alxatib and Pelletier, this framework suggests pragmatic maxims govern which interpretation is taken (although here the notions of overlap and underlap aren't based on supervaluationism and subvaluationism as they were in Alxatib and Pelletier, but rather the strong Kleene logic K3 and the paraconsistent LP). In this case, the suggestion is that participants begin applying 'blue' to the clear cases in a *strict* sense, and prefer to switch to a *tolerant* sense of 'green' (as borderline blue-green shades are tolerantly both) than to reinterpret 'blue' in a tolerant sense. Once the borderline cases have been reached, it is more informative to shift category to highlight the semantic change in the colour shades, even though a tolerant use of 'blue' would still be accurate.

In the second experiment, a statistically significant negative hysteresis is found, replicating the finding in the linguistic task in the first experiment. Participants agreeing to 'the square is yellow' (or disagreeing to 'the square is not yellow' etc…) would shift to disagreeing at an earlier point in the series to that at which they shift from disagreement to agreement when traveling in the reverse direction. The authors also report what they call a 'hump effect' when it comes to acceptance patterns for the conjunctive sentences and their conjuncts. The term is borrowed from Ripley (2011), who reports that participants are more likely to assent to contradictions when they describe borderline cases than clear cases, in particular in his case, more likely to assent to 'the circle both is and isn't near the square', but Egré et al really use the term to refer to the sort of result Alxatib and Pelletier report, that participants would assent to the conjunction whilst individually dissenting from the conjuncts. The test this, Egré et al aggregated the atomic data across all colour and order conditions, and interpolated the point in the colour spectrum where the individual conjuncts (e.g. the square is yellow, the square is orange) were equally acceptable (i.e. where the percentage of agreement was equally high), and compared it to the interpolated degree of acceptance to the conjunction (e.g. the square is yellow and orange). They report the difference of acceptance is statistically significant in every case, except for 'the square is yellow and not yellow', with an average difference of 12% for the yellow-orange series, and 21% for blue-green series, between the acceptance of the conjunction the acceptance of each conjunct. It is suggested that this result is a replication of Alxatib and Pelletier's:

> this result is exactly consistent with the finding of Alxatib and Pelletier (2011) in the case of static stimuli relative to the vague predicate "tall", and it lends support to the overlap hypothesis we used to account for negative hysteresis in the first place. (2013: 415)

Egré et al take this result to give further support to their view, based on the dual tolerant and strict interpretations of vague language. As also argued by Alxatib and Pelletier, Egré et al argue that to explain why participants would accept a conjunction more readily than the conjuncts, we need to be able to make use of an ambiguity, such that participants *tolerantly* accept a contradiction, but reject a strict interpretation of each conjunct. However, the authors don't explicitly use this result to reject the epistemic and supervaluationist views discussed after the results of the first experiment.

3ii. Criticising Egré et al

Egré et al have at least one strength that Alxatib and Pelletier lacked, being that they only consider there to be an important result in the data where it is statistically significant. However, there still reasons to be dissatisfied with their conclusions, on a practical and theoretical level.

The (ir)relevance of hysteresis

A large part of this study is dedicated to hysteresis effects, and what it might tell us about vagueness. The authors initially suggest a *positive* hysteresis that would support an overlap view of borderline cases (i.e. 'glutty' theories of vagueness) and suggest that a *negative* hysteresis supports an underlap view (i.e. 'gappy' theories of vagueness). This view is expressed elsewhere, e.g. in this Smith quote (which Egré et al endorse):

> Suppose we are walked along our Sorites series for F, and asked of each object in the series whether it is F, and then walked back the other way, and asked the same question of each object again. It is very likely that the point at which we stopped saying 'Yes' on the way out would be further along the series than the point at which we started saying 'Yes' on the way back. This behaviour might seem rather difficult to explain on the recursive truth gap view, but it is easily explained by the contextualist. (2008: 118)

However, following the first study, despite finding a negative hysteresis, they take the result to support the pluralist theory seen in Cobreros et al (2012), and suggest that participants represent the borderline regions in the series as an *overlap*: 'while plausible to us, the explanation proposed at the end of the previous section supposes that the negative hysteresis found in the linguistic task is a robust effect, and more over that subjects do indeed recognize an overlap region in each color set' (2013: 405). They address this explicitly later in the paper, ultimately suggesting that both positive and negative hysteresis can support an overlap view:

> …both for situations of [positive] hysteresis and for situations of [negative hysteresis], there is a range of cases that subjects equally call "blue" and "not blue", or "blue" and "green", depending on the direction of the transition. In that sense, the finding of [a negative hysteresis] remains entirely consistent with the characterization of vague predicates in terms of an "overlap". Furthermore, we have independent evidence for the overlap view, based on the fact that subjects agree to sentences of the form "the square is blue and green", or "the square is blue and not blue", in Experiment 2. (pp. 417)

This sounds right. In both cases of negative and positive hysteresis, the borderline range will be described in two different ways, depending on the direction of travel along the continuum, so it seems wrong that we should only take positive hysteresis to support an overlap view. However, whilst I endorse this point, it now puts a question to the purpose of investigating hysteresis here. After all, this shows that *whatever* result they found in the first experiment,

the authors would have had something consistent with their pluralist strict/tolerant theory. If it were a positive hysteresis, then they could have concluded that subjects were interpreting the vague words tolerantly in each direction along the continuum, supporting an overlap view. Finding a negative hysteresis, they concluded that subjects initially interpret the vague words *strictly*, and prefer to shift to tolerant interpretations of the alternative category upon reaching the borderline cases, which shows the borderline range can be tolerantly described as either. Even if no significant hysteresis was found, their view would remain consistent with the result; it could be argued that participants do interpret borderline ranges as an overlap, but over a range too small for statistical tests to pick up. This puts a serious question to the relevance of the results of the first experiment; if either a positive or negative hysteresis would have supported their preferred *overlap* view of vague predicates, and their strict/tolerant framework, then we haven't gained anything from discovering there was a negative hysteresis in the linguistic task.

The same can be said for any of the views considered following the first experiment. The supervaluationist theory of vagueness needn't predict a negative hysteresis. As noted, either hysteresis would show that speakers are willing to describe borderline blue/green cases as both blue and green under certain circumstances, so it isn't obvious why we should think a gap theory predicts an earlier shift rather than a late one. Egré et al offer the auxiliary assumption that speakers prefer to shift category to be informative of the change in semantic status (from blue to borderline), and prefer to use 'green' than distort the accurate use of 'blue' to that point. But that assumption is very speculative, and there's no reason a supervaluationist is committed to saying speakers would act that way. If a positive hysteresis were found, an equally plausible auxiliary assumption could be used, for instance that speakers simply prefer to continue to use 'blue' than apply 'green' to a borderline case, as it would be inaccurate; they save 'green' for when it can used to inform of the change of status to truly green shades. Again, there is no good reason why supervaluationism need predict anything in particular about behaviour in the context of a hysteresis. As both negative and positive hysteresis is plausibly consistent with the supervaluationist framework, the evidence of a negative hysteresis in the first experiment isn't relevant to the plausibility of a supervaluationist interpretation of vague predicates.

Similarly, Egré et al suggested an epistemicist could explain the negative hysteresis with the preference of errors of omission to commission. However, it ought to be noted that it isn't clear that this is the same kind of omission/commission distinction see in Bonini et al. In

Bonini et al, they could appeal to a clear omission of applying a predicate to a certain range (i.e. some heights near the borderline region) for fear of misapplying it and falsely committing to saying some heights are in the extension of 'tall' that are not. It's not clear this translates to a negative hysteresis; switching category early may allow the participant to avoid an error of commission with the first category, but means they will be committing an error of commission with the second category. Both negative and positive hysteresis will involve some error of omission and commission from the point of view of epistemicism. Either speakers switch category too early, to commit to falsely applying the second category, or they switch too late and commit to falsely apply the first category. As both negative and positive hysteresis means describing some range of shades as two categories, there is no clear reason that the epistemicist need be committed to one kind of hysteresis or the other; they can claim that either pattern of response shows a preference of for omission or commission, if they think there is such a bias.

This puts a serious challenge to the significance of the first experiment. Any of the three candidate theories are consistent with both a positive and negative hysteresis, so it's unclear what the role of investigating hysteresis is. There of course are reasons to be interested in hysteresis effects; one might simply be interested in how speakers are likely to behave in a 'forced-march' along a sorites series. But Egré et al make it clear they aim to investigate it 'in order to shed further light on the semantic status of borderline cases' (2013: 393). The results of the first experiment are weak evidence for any of the considered theories of vagueness, and don't add good evidence for their strict/tolerant overlap based interpretation; given any result was consistent with this view, there was no danger of a result that would go against their preferred overlap interpretation (except perhaps participants drawing a reliable sharp boundary in the same place in the series, but we can feel confident that such a result was very unlikely).

Ultimately, what is missing is a strong account of what *particular* empirical result in the hysteresis study would have been consistent with supervaluationism, epistemicism or the strict/tolerant framework. As I have stressed already, to really get something out of the experimental approach, we need a good account of what different theories would predict in a study. Egré et al do offer some ideas about this, but it now seems that their preferred theory would have been consistent with *any* hysteresis result, as would have been any of the other theories considered, and thus the result doesn't add good support for any of these theories.

One thought might be that the result is interesting if it *isn't* consistent with some other theories of vagueness. That would be interesting if it could be shown, but that isn't the argument made by the authors. Perhaps one defence to the worry is to point out that the strict/tolerant framework is the only theory of the three considered that is consistent with the negative hysteresis in the first study, *and* supported by the results of the second study. But again, as any result in the first study would have been consistent with the strict/tolerant framework, it is only the second experiment doing the work there, to which I now turn my attention.

The 'hump effect'

Whilst the role of hysteresis in the study is questionable, the second experiment contained a second factor, the assent to contradictions for borderline cases. Their primary result in this case was that the degree of acceptance of a contradictory conjunction (the square is yellow and not yellow) was higher than the acceptance of each conjunct, based on the point in the series where each was equally applicable. The idea was that this replicates the result in Alxatib and Pelletier, that speakers will assent to the conjunction more readily than either conjunct, and thus supports a similar pluralistic theory. As already discussed, I'm critical of the notion that this was a robust finding in Alxatib and Pelletier (with only 11 of 76 participant responding in that way), but Egré et al can at least claim that there is statistical support in this case; the degree of acceptance is statistically significantly higher for the conjunction.

Whilst I grant that this is an interesting result, it cannot be claimed that it is a replication of the result in Alxatib and Pelletier. There the idea was that speakers *dissent* from the conjuncts all together whilst accepting the conclusion. What Egré et al find is that the degree of acceptance of the conjunction is higher than that of the conjuncts at an interpolated point in the series where the degrees of acceptance are equal for each conjunct. All the authors tell us beyond this is that there was a difference of 12% for the yellow-orange series, and 21% for the blue-green series. The problem is that these differences in acceptance between the conjuncts and the conjunction are perfectly consistent with moderate or even quite high degrees of acceptance for the conjuncts. Say for instance that there was a 50% degree of acceptance for both 'the square is yellow' and 'the square is orange, and 62% degree of acceptance for 'the square is yellow and orange'. Even if that is a statistically significant difference, it isn't the kind of pattern that supports the *strict/tolerant* framework Egré et al are

defending. What is needed is a rejection of each conjunct whilst an acceptance of the conjunction. Whilst in this case there is a greater chance of participants accepting the conjunction than the conjuncts, there is clearly still quite high levels of assent to the conjuncts[4]. Thus, it is not convincing there is good evidence of the *strict/tolerant* theory here.

Overall then, Egré et al are not justified in concluding that there is evidence for the *strict/tolerant* framework here. The results of the first experiment cannot really be said to support any particular theory of vagueness, and the analysis of the second experiment doesn't show what would be required for good evidence of that theory. Whilst I endorse the role statistics play in this paper, once again there is not a good enough account of what kind of results would support what kind of theory, and we cannot accept the author's conclusions.

4. Conclusion

I hope that this discussion has outlined some of the difficulties facing the experimental approach. Whilst I do think studies like the above have the potential to be genuinely interesting and useful, I do not think we'd be correct to accept any of the author's main conclusions. I have criticised the studies for many reasons, including the design, methodologies, and theoretical set up. But in each case, we can see that the authors lack a satisfactory account of what relevant theory of vagueness might be said to predict in the study, or what kind of results are consistent with particular theories of vagueness. Why should a supervaluationist theory predict participants assent to '#2 is neither tall nor not tall', when this proposition comes out as necessarily false? Why would a 'gappy' theory of vagueness predict a negative hysteresis?

For answers to these questions we need a stronger theoretical basis, that offers an account how we can answer those kinds of questions based on the semantic and logical features of a theory of vagueness. It shouldn't be taken as obvious how indeterminacy, truth value gluts, or degrees of truth relate to beliefs and action. What should a supervaluationist say about attitudes towards 'x is red' when that is indeterminately true? Or what should degree theorist say about attitudes towards it when it is 0.6 true? In order to make progress with the experimental approach, we require a better theory of how these non-classical values should be taken to relate to these cognitive notions. This is where I turn my attention for the

[4] To go further we'd need it to be shown that the levels of assent to each conjunct were also quite low, and the levels assent to the conjunction was quite high, but Egré et al don't give us further details of these data points.

remainder of the thesis. First, I will aim to bridge the gap between a logico-semantic theory and a *cognitive* theory, and then I will offer accounts of what kind of cognitive theories supervaluationism and degree theories are committed to.

Chapter 4 Theories of Vagueness and Theories of Mind

In the previous chapter I argued that philosophers conducting experiments intended to shed light on the correct theory of vagueness were lacking in many respects, including design, method, and most importantly, their theoretical underpinning. They were seen to lack a *good* account of what kinds of results are well explained by theories of vagueness. In order to progress with the experimental approach we need to develop a more sophisticated view about the connection between a theory of vagueness and speaker behaviour in the experimental context, in order to speaker with more authority when we aim to connect a pattern in the data with a particular logico-semantic model. If we are to be successful in this task, we first need an account of the connection between a semantic and logical model on the one hand, its consequences for belief and decision making on the other. Examining such a connection is the aim of this chapter.

How can we move from a logical or semantic fact about say *borderline cases* to a fact about how subjects mentally represent borderline cases, and the decisions they will make in an empirical study? Is it right to make such a move at all; need the claim that a borderline case is true to degree 0.6 commit us to a claim that subjects will be in any particular sort of psychological state when making judgements about it? To make progress in the experimental approach we need to answer these questions, and to do so we must determine what the connection is between theories of vagueness, and theories about *rational belief and decision making*, what I will call *theory of mind.* What sort of restrictions, if any, do logics and semantic theories place on associated cognitive theories? Does endorsing a particular logic and semantics place constraints on what sort of attitudes we think subjects should or do have towards propositions with different semantic statuses? We cannot make assumptions about such a connection; many philosophers are happy to develop theories about vagueness, or the liar paradox, or some other phenomenon, which require non-classical approaches, without considering what impact this would have on theories about what attitudes subjects have to propositions affected by these phenomena. Similarly, some philosophers are happy to develop psychological theories about these sort phenomena, whilst claiming to remain neutral on the question of logic and semantics.

Here I will explore the possible connections between logical and semantic theories, and theory of mind, and defend the suggestion that the two do place some sort of restriction on each other. This is the first important step required by the experimental approach, lacking in current studies. This claim is grounded in two smaller claims: 1. Belief aims at truth 2. Logic has a normative role for belief. Given 1, which I take not to be a controversial claim, what a semantic theory labels as 'truth' is significant for cognitive theories. If truth behaves differently in two different semantic theories, we should expect this to result in some differences in cognitive theories associated with them. 2 is not so readily accepted, or at least, there doesn't seem to be a readily accepted way to characterise the normative role of logic. However, once we have characterised it, we can see that changes in the logic result in different norms for belief, resulting in changes in the theory of mind.

Whilst others have accepted that connections such as those between belief and truth and logic mean that logico-semantic theories will come with consequences for cognitive theories, it's not clear how strong these consequences are. I will begin by largely endorsing the way Robbie Williams (2014) sets up the notion of theory of mind, and how it can relate to logico-semantic theories, however I will differ from him in how strong he takes the relation to be. I distinguish between three different strengths of this relationship. On the weakest reading, your logico-semantic position places very little constraints on what cognitive theory is compatible. On the strongest reading, accepting a logic and semantics comes with a very particular constraint on your theory of mind; on this reading if you need to understand these constraints to be able to understand the theory at all. I will defend something in between these two; logico-semantic theories do place constraints on associated theory of mind, but are not so strong as to always delineate a particular theory of mind, there may be more than one compatible option available in some cases.

The upshot of this for the experimental approach is that it *is* right to take the logico-semantic framework given a theory of vagueness to commit us to *some* claims about the decisions subjects should make in an experimental study. However, it is not the case that a theory of vagueness is committed to particular ways of responding; there is more than one way to interpret the cognitive implications of the logico-semantic model.

1. Logic and Theory of Mind

My aim here is to explore, and defend a version of, the connection between logico-semantic theories and theories of mind. What is meant by 'theory of mind', or a 'cognitive theory'? I mean a cognitive theory about rational belief, desires and decision making (I will go into more detail shortly). The most relevant part of the theory of mind for the purposes of this paper is rational belief and reasoning. The question I want to ask is how might the features of a semantics and logic place restrictions on our associated theories about beliefs and attitudes? My primary interest is of course with the theories of vagueness, and I will be using supervaluational models as a case study, however I intend for the result to be generalizable.

It appears that some philosophers are not convinced there is a close connection between a semantics or logic, and theories of mind. For example, Rosanna Keefe claims that we shouldn't expect logical and semantic parts of a theory of vagueness to tell us much about a related cognitive theory:

> I would argue that we should not expect one's theory of vagueness to deliver an account of what psychological states we have (or even should have) in the face of borderline cases. Adopting some particular theory should allow us to recognise that subjects are alert to the difference between definite Fs and borderline Fs, but should be compatible with a range of attitudes to borderline cases… (2016: 3789)

In a mirroring example, Stephen Schiffer develops a psychological model of vagueness, claiming that what constitutes vagueness is his notion of partial belief (2000: 229-231). He then claims that his view is 'evidently neutral with respect to the semantic, logical and metaphysical question' of vagueness (2000: 231). Here Schiffer has expounded a cognitive theory, but suggests it can be kept quite distanced from logic and semantic theories.

I will be arguing that there is a stronger connection between our logico-semantic theory of vagueness and cognitive theory than it seems Keefe and Schiffer are suggesting here. On the opposite end of the scale would be the suggestion that there is a very strong connection between logico-semantic theories and cognitive theories. It might be thought that an example of such a position is that of Robbie Williams, seen in the following quote (where *T* is some non-classical theory):

> The theorists themselves, in accepting *T*, adopt some stance or other toward the propositions that *T* describes as indeterminate, whether that stance is belief, disbelief, half-confidence, suspension of belief, or something else entirely. In doing so, they accord with some accounts of the cognitive role of indeterminacy and violate others. That is why I claimed above that in accepting *T*, theorists already implicitly take on commitments in the non-classical theory of mind. (2014: 383)

Williams is suggesting that in just the acceptance of some non-classical theory, one is already taking a stance in the theory of mind, about what attitudes agents takes to sentences with non-classical truth values. Williams argues that those proposing non-classical logics and frameworks should be developing associated non-classical cognitive theories, his key reason being that a non-classical theory *will* have some commitments to explain in the non-classical theory of mind.

I think Williams is going in the right direction. I do agree that accepting a non-classical theory will come with *some* commitments in the theory of mind, however I worry that Williams is taking them to be too strong here; he suggests in the above quote that accepting a theory means adopting a particular cognitive theory. I will argue for a position in between the strengths of Keefe and Schiffer, and Williams. I think there are restrictions placed on the cognitive theory by logic-semantic theories, however they are not as strong as Williams supposes here: they may restrict cognitive theories, but may not necessarily commit you to a particular theory.

1i. 'Logic' and 'theory of mind'

What do I mean by 'logic', 'semantics' and 'theory of mind'? I take semantic theories to give truth conditions for truth bearers. This will involve an account of what possible semantic statuses there are, under what conditions atomic truth bearers receive these statuses, the semantics of logical operators, and the truth conditions of complex truth bearers. I take logic to be the specification of a consequence relation on this set of truth bearers, and consequence to necessarily preserve truth. The logic tells us what truth bearers are tautologically true, which are contradictory, and what inferences are valid.

By 'theory of mind', I am roughly following what Williams means (2014: section 1). The theory of mind is our theory of rational belief[5], desires and action, and how these three inter-relate. It accounts for what sort of attitudes agents do (perhaps should in cases, I will discuss this shortly) have, how these attitudes change given new evidence, and how this interacts with decision making; with a notion of desires and choosing actions. Williams suggests that one way of capturing this sort of theory of mind is with the following claims:

[5] One might question the choice of '*rational* belief' here, for the reason that it might not look clear whether what a theory of vagueness is committed to saying about rational belief commits them to saying anything about what beliefs speakers *actually* have, i.e. when we test them in an experimental context. We can avoid concern here with the assumption that speakers are rational, and thus will take up rational beliefs.

Probabilism: Ideally rational beliefs come in degrees and meet the constraints of probability theory

Conditionalising: The rational way for an agent with beliefs *b* to respond to receiving total evidence *E* is by moving to the belief state *b'* that results from conditionalizing *b* on *E*

Utilities: The rational belief-desire states for an agent to be in are those that are represented by probability-utility pairs meeting the constraints of Jeffreys' (1983) decision theory

Maximising: The rational act to choose, out a given set of options, is the one that has maximum utility

This gives us a good idea of what to expect from a theory of mind. We have an account of what rational belief states are, and of belief update. We also have an account of how these beliefs relate to decision making; through assessments of utilities based on belief and desires, and rational action. For the purposes of this chapter I will be primarily interested in rational belief, and how that might be constrained by a logic and semantics. An important aspect of this is identifying the rational attitude to take to different semantic statuses. This is something that we can gather from the theory of mind above. *Probabilism* is committed to classical probability theory, according to which a probabilistic belief in some proposition P is a function of how likely an agent thinks it is true plus how likely an agent thinks it is false; credence in P = (1 x how likely P has value 1) + (0 x how likely P has value 0). Where we are not certain of whether P is true, this will give a degree of belief between 1 and 0. Where we are certain of P's truth value, we get a clear picture of what attitude is rational to take: belief (credence 1) where P is true, disbelief (credence 0) where P is false. This is of course only suitable for a classical framework, as *Probabilism* commits agents to the constraints of classical probability theory, which assumes bivalence. If we were to adopt a non-classical theory, it would tell us nothing about the attitudes we should have to non-classical truth values. Similarly, classical probability theory assumes classical logic; as all classical tautologies have probability 1, and classical contradictions have probability 0, but non-classicists may well want to deny either that all classical tautologies are true, or that all classical contradictions are false.

My investigation here is into what way accepting a logic or semantic framework commits you to particular features of a theory of mind, like that above. Where these frameworks are non-classical, like those that I'm interested in, the theory of mind considered thus far would need

adapting or generalising. Williams gives us one way of doing so, suggesting the following generalisations of *Probabilism* (where *b*(A) reads as 'belief in A'):

Belief-Truth: A rational belief state *b* is *probabilistic1* iff for each A, *b*(A) is a suitable weighted average of the truth values *A* takes at worlds

Belief-Logic: A rational belief state *b* is *probabilistic2* iff *b*(A) is 1 where A is a tautology, 0 where A is a contradiction, $b(A) \leq b(B)$ where A entails B, $b(A \vee B) + b(A \& B) = b(A) + b(B)$

These give us a way of generalising the notion of rational belief in the above theory of mind. We get a different cognitive framework depending on what logic and semantics we plug into it. On a classical model, *Belief-Truth* and *Belief-Logic* would give us classical probabilistic beliefs. Propositions are either true or false, and our probabilistic belief in a proposition is a function of how likely we think it is true, and how likely we think it is false, a weighted average of its truth value at worlds as Williams puts it. We fully believe propositions we are certain are true, and disbelieve those we are certain are false, fully believe classical tautologies and disbelieve classical contradictions.

We can also read 'truth values' in *Belief-Truth* non-classically, and our probabilistic beliefs become a weighted average of non-classical truth values. Similarly, we can take the tautologies, contradictions and entailment referred to in *Belief-Logic* to be non-classical. For instance, take a paracomplete theory, with the truth values true, indeterminate, and false. We could read these as represented by the set {1, 0.5, 0}, and then plug them into *Belief-Truth.* Probabilistic belief in a proposition would then come out as a function of how likely it is that that proposition takes truth value 1, 0.5, and 0. Where we are certain of the proposition's truth value, we would fully believe it if it's true, disbelieve it if it's false, and take a half credence if it's indeterminate. The tautologies and entailment mentioned in *Belief-Logic,* would constrain belief based on the tautologies of this three-valued logic and its consequence relation; for instance, this being a paracomplete logic, LEM would no longer be a tautology and require full belief. We can get similar results by plugging in other non-classical logics and semantics, such as degree logics and paraconsistent theories.

This generalisation can be motivated in the same way classical probabilism is, based on Dutch book and accuracy domination arguments (see Paris 2001, Williams 2012). Given this, we can go on and give generalised notions of belief update, and the connections between

belief, desire and action, as Williams does (2014: 388-9). This does give one way to approach non-classical theory of mind, but it's not clear it's one we should accept. For one thing, we can question whether beliefs do or can always match non-classical truth values. Consider again our paracomplete approach. Not all theorists may want to accept the framing of indeterminacy as 0.5 true, but rather want to claim indeterminacy is the lack of a truth value, just 'neither true nor false'. This won't sit so comfortably with *Belief-Truth*. We can't take belief to be a weighted average of truth values at worlds, as there is now no numerical truth value for 'indeterminate' to input into the equation. They might also reasonably worry whether a half-credence is the right attitude to take towards indeterminacy. Consider also a degree logic and semantics; a theorist may want to deny that credence can or does accurately match infinitely many truth values. As I will discuss in more detail in chapter 6, there are alternatives on which belief isn't modelled as a precise value matching a precise degree of truth.

Williams does point out he doesn't want to claim that all non-classical theories can fit into the space of non-classical mind delineated here, but degree and paracomplete options are of the type that do fit for him. My main concern can be directed with his claim that 'we need to associate each alethic status with a "cognitive load"- a numerical representation of the level of confidence that God would have in A, when A has the status in question. We call this the truth value of A (at the relevant world w) …' (2014: 386-7). I'm not sure theorists need always accept that truth-values straightforwardly pick out the right level of confidence to have in a proposition with that truth-value. I do endorse that the only rational level of confidence for truth is full confidence (credence 1), and for falsity is no confidence (credence 0). But when it comes to non-classical semantic statuses, it isn't as obvious what sort of confidence it is appropriate for agents to have. As seen with the paracomplete example, there may not be a truth-value to associate confidence with. With the degree case, whilst an omniscient god *can* match confidence to infinitely many truth values when propositions have those values, a degree theorist might claim that we're more interested in modelling the sort of belief that non-omniscient humans can have, and claim that rational belief needn't match up with finely grained degrees of truth, but that rational beliefs may be something softer, less precise.

This is an important issue on which I will depart from Williams. I agree that the way in which semantic theories identify truth values is relevant to their cognitive role, and this means that the relationship between them and theory of mind is stronger than perhaps supposed by Keefe

and Schiffer in those above quotes. However, I don't think it's as straightforward as identifying each truth value as representative of the level of confidence it is rational to have in it. We'll see this in more detail in the following section.

2. Three different strengths of commitment

Hopefully what I mean by theory of mind is now clearer, as well as what I mean when I say I'm interested in the connection between a logico-semantic theory and rational belief. Now I turn to the central issue. What sort of constraints does accepting a particular logic and semantics place on the associated theory of mind? I see three different possible versions of the commitment theories may have in theory of mind:

Weak commitment: a logic and semantics does not place any particular restrictions on theory of mind. Accepting a logic and semantics needn't tell us anything about the rational attitude to take towards different alethic statues, or other features of rational belief and decision making

Medium commitment: a logic and semantics place *some* restrictions on the theory of mind. Particular features of the semantics and logic entail that some features of a theory of mind are ruled out, and but there may remain more than one possibility in some cases

Strong commitment: a logic and semantics are committed to a *particular* theory of mind. That this is part of what distinguishes a semantics or logic; on this view the logico-semantic model is not entirely separate from the theory of mind. We need to understand the commitments in the theory of mind to fully understand the logic and semantics

Weak commitment is the sort of view I've attributed thus far to Keefe and Schiffer, based on those quotes. It's one on which philosophers can continue to develop semantic and logical theories, without needing to worry about consequences for the theory of mind. A proponent of this view may accept that if they take a non-classical theory, the classical theory of mind isn't entirely suitable, at least in as far as it doesn't offer any advice on how deal with non-classical statuses or logical consequence, but think it's open beyond that. *Strong commitment* is the sort of view I think can be attributed Williams. In that first quote he claims that accepting a non-classical theory of some phenomena means already accepting some stance towards that phenomena. He also suggests elsewhere that in order to really understand a non-

classical theory, we need to know what accepting it would mean for the theory of mind (2014: 383).

Medium commitment is the sort of the view I want to defend. It suggests, pace *weak commitment*, that a logic and semantics may well place some restrictions on a theory of mind; we shouldn't expect a non-classical theory to be consistent with taking any sort of attitude to non-classical semantic statuses. But also, pace *strong commitment*, we shouldn't expect a non-classical theory to be consistent with only *one* theory of mind. The trichotomy of positions I've laid here might not be the end of the story. *Strong* and *weak commitment* give two extremes, and *medium commitment* broadly covers almost anything in between. There may well be different approaches that would fit into my notion of *medium commitment*, but for now, as I'm aiming to show that *weak* and *strong* are wrong, I'll stick with it.

First, I'll argue that the correct position must be stronger than *weak commitment*, in the following chapters I will argue against *strong* commitment, when considering the possible options for supervaluationism and degree theories. This is based on plausible connections between truth, logic and belief, which have already been alluded to. I suggest that the notion of 'truth' has a cognitively significant role to play when it is identified in a semantics, partly based on the suggestion that belief *aims* at truth. I also suggest that logic is normative for belief, so we should expect changes in logic to come with changes for rational belief.

2i. The cognitive role of truth and logic

My reasons for rejecting *weak commitment* are based on the claim that truth and logic come with an important cognitive role. For truth, this is partly revealed through the commonly accepted platitude that belief 'aims' at truth. Indeed, it is hard to make sense of what beliefs are if we deny that they are aimed at truth. It is clear that a belief is correct if it is true, and that it is somehow incorrect if it is false. This shows that the notion of truth intuitively has a cognitive role to play. When 'truth' and 'falsity' are identified in a semantic theory, we are identifying the structures, or properties of propositions, which we ought to respectively aim belief and disbelief at. If 'truth' were identified in a semantic theory, but it didn't have this cognitive role, it would be somewhat mysterious what its role in the theory is, even what justifies giving it the name 'truth'. Theories that differ in their suggested features of truth, but not in their associated theory of mind, I will argue, may not be different in a substantive way.

It is for this sort of reason that we should expect the features of truth in a semantic theory to come with cognitive significance. I will shortly try to make this point more explicit with a case study on supervaluational frameworks, and different possible semantics that can be delineated. There I will agree with Williams that for two semantics to be genuinely substantively distinct, they must come with different theories of mind (2014: 393).

The cognitive role of *logic* is slightly more complicated. Logic has a cognitive role insofar as it is normative for belief. The suggestion is that two different logics norm belief in different ways, and thus change the features of rational belief in the theory of mind. What the normative role of logic is or may be is a complicated question, worthy of its own thesis, and there isn't space to give the question a full treatment here. But I can outline some of the issues that face characterising a normative role for logic, consider how to meet them, and move forward with a reasonable enough looking way of handling it.

It is plausible that logic is epistemically normative in some sense. It seems that an agent which accepts that 'grass is green' and 'the sky is blue' ought to also accept their conjunction, and that it would be rational to do so under some circumstances. For instance, if I asked such an agent, 'ah so you think that grass is green and the sky is blue?', a rational answer would surely be 'yes'. It certainly seems that they would be making a mistake of some kind if they were to say anything else. I think this is the case because our beliefs aim at truth. As the truth of A and B necessitate the truth of 'A & B' (assuming classical logic for now- this will be addressed later), it is irrational for an agent to accept the former two but refuse to accept the latter. The difficulty is trying to pin down exactly what sort of norms are placed by logic, and in what way they constrain belief and rationality.

How do we go about finding a characterisation of the normative role of logic? We need a principle that would take us from a fact about logic to constraints on belief. A common first approach is something like the following:

N1: Where $A \vDash B$, if you believe A, then you ought to believe B

This captures something that we want from a characterisation of the epistemic norms of logic, alluded to above. If an agent believes A, and A entails B, then in some sense they should also believe B. It would be irrational to disbelieve B, as it would be inconsistent with your belief in A. *N1* delivers this result, but is not a satisfactory principle. Gilbert Harman (1986)

famously presents challenges to a connection between logic and rationality, of which *N1* falls foul.

First, it seems like *N1* is straightforwardly false; if B were an absurdity, then it's not the case you *ought* to believe B, even if your belief in A entails it (it seems here you should instead revise your belief in A). Second, *N1* is *demanding*. According to it we *ought* to believe any logical consequence of our beliefs. But it's not possible for us to recognise every logical consequence of our beliefs, there will be infinitely many of them, and an agent certainly shouldn't be considered irrational for failing to do so. Third, *N1* is *cluttering:* according to it, we ought to believe any logical consequence of our beliefs, but for any belief, there will be an infinite amount of irrelevant logical consequences. From A, it follows that A v A, A v A v A… Given *N1,* an agent would be irrational for failing to believe all of these consequences, but that is clearly false.

N1 then, is too strong. One improvement is *N2:*

N2: Where A ⊨ B, you ought to see to it that where you believe A, you also believe B

N2 has widened the scope of the *ought,* which allows us to get around the first problem *N1* faced. If you believe A, it's not the case you ought to believe B, even if B is absurd. You just ought to see to it that if you believe A, you also believe B. So, if you believe A, which entails an absurdity B, it's the case you ought to believe B or reject A. However, *N2* is still demanding and cluttering. According to this principle, you should either reject axioms of Peano arithmetic, or also believe every theorem, an impossible task. Similarly, you ought to either give up your belief in A, or believe all of A v A, A v A v A… *N2* is still too strong.

2ii. MacFarlane's principle

John MacFarlane (2004 (unpublished)), who also rejects the two above approaches, suggests an alternative:

N3: Where A ⊨ B, you ought to see to it that if you believe A, you don't disbelieve B

This changes the obligation from one of belief to 'not disbelief'. This principle is not demanding or cluttering, as we are only obliged to ensure we don't believe a proposition whilst disbelieving its logical consequence. For instance, we can accept the axioms of Peano arithmetic without having to believe all the theorems, insofar as we don't *disbelieve* them.

And we needn't take up the infinitely many disjunctions that follow from any of our beliefs or reject every belief, as long as we don't disbelieve any of those disjunctions.

However, *N3* is ultimately too weak. If this were the right principle, then someone who believed A, believed B, but refused to take up a stance on A & B wouldn't be making any sort of mistake, even if they recognised the entailment. As long as this agent didn't disbelieve the conjunction of A and B, then she is acting in rationally by the lights of *N3*. Whilst *N3* is weak enough to avoid problems of demandingness and clutter, it isn't strong enough to capture something important about a normative role of logic. Call this the problem of obtuseness. As pointed out at the outset of this discussion, someone who accepts some premises, notices a valid entailment from them to some conclusion, but *refuses* to accept that conclusion seems in some way irrational. *N3* cannot capture the irrationality here.

To deal with this issue, MacFarlane suggests that we also acknowledge that logical entailment gives (defeasible) reasons to believe: where A ⊨ B, one ought to see that where they believe A they have a reason to believe B. The agent above who believes A, and B, but refuses to take a stance on A & B, must now see that the belief in A and B and the entailment of the conjunction gives at least a *reason* to believe their conjunction. MacFarlane takes this to solve the problem of obtuseness, but we shouldn't find it satisfying. The idea was that there is something *irrational* about refusing to accept the clear logical consequence of your beliefs. We now have at least a reason to believe B when it follows from A, but one can hardly be thought *irrational* for refusing to believe B on account of a single defeasible reason.

Consider someone who fully believes that 'grass is green', and 'the sky is blue', and acknowledges that those together entail 'grass is green and the sky is blue', but refuses to take any attitude to the conjunction. This person is clearly making a mistake, they are being irrational. But this irrationality must go further than simply failing to do something for which there is a reason to do. As this reason is *pro tanto*, an agent could very rationally fail to believe or disbelieve 'the grass is green and the sky is blue' if there were some reason to do so, which outweighed the reason to believe provided by the logic. Someone who recognises the validity of conjunction-introduction to two of their beliefs, and refuses to take an attitude to the conjunction is making a more fundamental error. They recognise that some propositions they believe guarantee the truth of another proposition, but refuse to take an attitude towards it. It is like claiming 'P is true but I don't believe it or disbelieve it'. As

discussed, the rational attitude to take to true propositions is to believe them. Recognising a proposition as true but refusing to take an attitude towards it is irrational. This is what lies as the heart of the irrationality of the obtuseness problem, and acknowledging that entailment gives reasons to believe does not satisfactorily capture it. It is not irrational to fail to map belief to consequence because of a *pro tanto* reason, it is because it is irrational to recognise a guaranteed truth and simultaneously fail to believe it.

Thus, I don't think *N3* gives us the whole story about the normative role of logic. I think the principle is probably true, and that there are reasons to believe propositions entailed by our current beliefs, but *N3* is too weak to capture the irrationality in the obtuseness case. A stronger principle can capture what it gets right and more. We can take to step further and say not just that we ought not disbelieve, but we do have obligation to believe in these cases.

2iii. Field's principle

Field (2009) gets closer to what we want. He suggests the following account of the epistemic norms of logic:

N4: Where it is obvious A ⊨ B, you ought to see to it that if you believe A then you believe B, where A and B are in question

A key improvement seen here is that there are only norms on belief when the relevant propositions are *in question*. So you only ought to be *disposed* to believe the logical consequences of your belief. This avoids being too demanding or cluttering. For instance, one isn't obliged to believe every theorem of Peano arithmetic or deny the axioms; it is only when an obvious theorem is in question that the constraint kicks in. Similarly, one isn't obliged to believe all of the infinitely many disjunctions that follow from our beliefs, or reject those beliefs; rationality only requires that one is *disposed* to believe any of these disjunctions when one and the proposition that entails it are in question.

Importantly, *N4* also gives a more satisfying response to the obtuseness problem. An agent who believed A and B, but refused to accept the conjunction of the two *is* making a mistake, they are being irrational, they *ought* to believe the conjunction (assuming that adjunction is an obviousness entailment, which doesn't too far-fetched, but more on this shortly).

However, I am suspicious of the idea that only *obvious* entailments place epistemic norms. For one thing, it can deliver some counter-intuitive results. On one hand, *N4* can come out

too weak. For instance, I know that Fermat's last theorem is true, but it certainly less than obvious based on it's very complicated proof. Given the complicated but successful proof, it should be irrational to accept the axioms of number theory and deny Fermat's last theorem. However, it's not clear that *N4* gives this result; as the theorem is not *obvious* by anyone's lights, *N4* presumably does not place any norms on our attitudes towards it. On the other hand, I worry it could be too strong. It will deem people irrational for failing to see obvious entailments, even if they don't *know/accept* they're valid ones. I am not convinced people should be deemed irrational for failing to map their beliefs to entailments they don't know/accept. For instance, if adjunction is deemed to be an obvious entailment, then someone is irrational for failing to map their beliefs over it, by the lights of *N4*. But consider an agent who accepts subvaluational logic (e.g. Hyde 1997), on which adjunction is not valid. It seems wrong that such an agent would be *irrational* for not using adjunction, when they don't accept it's a valid entailment, and may have good reasons for doing so.

But whether this agent is irrational hangs on whether adjunction is deemed obvious, which raises the most important question for *N4*; which entailments are meant to be the obvious ones, and who decides they're obvious. Is adjunction obvious, even if some logicians have good reasons for rejecting it? Is disjunction-introduction obvious, when it surprises some undergraduate students on their first logic course? I have reservations about whether a satisfying answer can be given here.

The worry connects to a general problem for characterising the normative role of logic. How are we to interpret the consequence relation referred to in the principles? Which logic exactly is meant to be placing norms and constraining our beliefs? Is it the 'correct logic', or the logic one adheres to? I don't think *N4* gives us the best answer to this. Imagine two consequence relations, $\vDash_1$ and $\vDash_2$, with the following properties:

$A \vDash_1 B, A \nvDash_1 C$

$A \nvDash_2 B, A \vDash_2 C$

Suppose myself and a friend both accept A, except I am a proponent of $\vDash_1$, and she is a proponent of $\vDash_2$. Although we both accept A, I go on to also accept B but not accept C, whereas my friend accepts C, but not B. How is one to decide when and when not my friend and I are being rational. If norms are of the form suggested in *N4*, it depends on when the

entailments are *obvious*- but again, obvious to who? And obvious entailments of *which* consequence relation?

Suppose that $\vDash_2$ is the correct logic, but we are both logicians with good reasons for accepting our chosen logics. Am I irrational for failing to recognise an obvious entailment of the correct logic, and don't believe C? Or would I only be irrational for failing to recognise obvious entailments of the logic I acknowledge? Does *N4* mean, whenever it is obvious that A ⊨ B… where '⊨' refers to the correct logic, i.e. $\vDash_2$? Or obvious entailments of the reading you accept ($\vDash_1$ in this case)?

Field suggests that *N4* can be read either way (2009: 262). When it comes to evaluating someone's epistemic practise, we can either assess by the lights of obvious entailments to *us,* or obvious entailments to the person we're evaluating, and that we keep it this way, sometimes doing the former, sometimes doing the latter. This also allows for some sort of plurality to norms governing people's beliefs. The way people 'ought' to modify their beliefs varies depending on which logic they're being assessed with, the one they adhere to, the one I/another judge adheres to. This is written with Field's expressivist view in mind, according to which all there is to 'accepting a logic' is following the epistemic norms it places (see 2009b). There is no particular 'correct logic', whether someone is being rational depends on which epistemic norms we choose to evaluate them with.

I'm not sure how satisfying this solution is. As it stands *N4* means that logic only places norms where entailments are obvious, but we have no useful account of how to determine which entailments are obvious ones. This adds unnecessary complications for my purposes, so I'd prefer to use a characterisation without the notion of obviousness here.

There is clearly plenty to discuss on this topic, and I don't intend to try settle it. It has been useful to examine the challenges that face a characterisation of the normative role of logic, and some ways of attempting to handle it. To progress with the main task at hand, I will use the following characterisation:

N5: Where A ⊨ B, you ought to see to it that where you believe A, you are disposed to believe B

This is the same as Field's *N4*, but with the notion of obviousness dropped out. It still shares the merits of *N4*. It is not demanding or cluttering. Agents can accept the axioms of Peano arithmetic without being obliged to believe every theorem, it is only if one is *question* that

they ought to believe it. An agent also isn't required to believe the infinitely many disjunctions that follow from each of their beliefs, it is only if the relevant belief and disjunction that it entails are in question that they are obliged to.

Like *N4*, *N5* also gives a much more satisfactory solution to the obtuseness problem than MacFarlane's suggestion, *N3*. Anyone who acknowledges their belief A entails another belief B but refuses to take a stance on it is making a mistake. And this mistake is more than just not doing something there was a defeasible reason for, there is an epistemic obligation to believe B or reject A that follows from the entailment from the latter to the former.

So *N5* gives us the key results we have thus far looked for. Its drawback is that it means that an agent *ought* to accept a logical consequence of one of her beliefs when it is in question, regardless of her attitude towards the relevant entailment; even if she isn't aware of it, or has well developed reasons to reject it, for instance. But that result only follows insofar as the entailment is a *true* one, i.e. an entailment of the correct logic. And if that is the case, then there is a reason that an agent ought to be disposed to believe the logical consequences of their beliefs; their belief would better track the truth if they did. If we are working in a context where we take it that there is no correct logic, then we can take the sequent in *N5* to be representative of how any particular logic constrains belief.

N5 gives us a suitable characterisation to continue with. It was initially plausible that, by modelling truth preservation, logic constrained belief in some way, and *N5* can give us a palatable way of capturing this: agents ought to be disposed to believe the logical consequence of their beliefs, where they are in question. This isn't too demanding, cluttering yet doesn't allow for logical obtuseness as *N3* did. I do not think *N5* is ultimately the correct characterisation of the normative role of logic, but the purpose of this thesis is to examine how theories of vagueness, as descriptive models, connect to theories of belief, and *N5* is close enough for now.

Before moving on, slight complications must be added to the account, which are suggested by Field (2009: 255). First, belief is not an all or nothing thing, so we must adjust the principle to allow for varying credence. The current principle is that we ought to see to it that we're disposed to believe B where we believe A. The general idea is that we don't have a belief in the premise without being disposed to also believe its logical consequence. When credences are introduced, the general idea becomes that we ought to be at least as confident in a logical consequence of one of our beliefs as we are in that belief. So where A entails B, we ought to

be disposed to see to it that our credence in B is greater than or equal to our belief in our belief in A (where b(A) reads as credence in A):

N5+: Where A ⊨ B, you ought to be disposed to see to it that b(B) is greater than or equal to b(A)

Also, it will be useful to adjust the principle such that we allow multiple premise arguments. In general, the idea would be that where a set of premises entail B, you ought to be see to it that you're disposed to believe B if you believe each member of the set. But we also need to incorporate degrees of credence into this picture. The result should be that we ought to (be disposed to) be more confident in a logical consequence of a set of beliefs the more confident we are of that set of beliefs, and ought to be disposed to fully believe any logical consequence of a set of beliefs each of which we fully believe, that we ought to be disposed to fully believe a logical tautology. The following principle gives these results:

N5:* Where A*1*, …, A*n* ⊨ B, you ought to be disposed to see to it that b(B) is greater than or equal to b(A*1*) +…+ b(A*n*) – (*n*-1)

Where *n*=1, credence in B ought to match credence in A (the same result from *N5+*). Where credence is 1 for all premises, credence in B ought to be 1. Where *n*=0 (i.e. ⊨ B), B is a tautology and credence ought to be 1. The upshot of this principle is that the more uncertain you are of the premises of the entailment, the less certain you ought to be of the logical consequent, and the more premises you add of which you are uncertain, the less certain you ought to be of the logical consequent. The idea is that the overall uncertainty in the set builds up and trickles through to the logical consequence of that set. This is the desired result: the more conditions required for an outcome, i.e. for Aston Villa to win the Champions League, each of which I am not certain will obtain, the less likely I should think that outcome is. If they have to win 2 games, which I am 90% confident they will win, then I might feel quite confident they will win the Champions League. If they have to win 5 games, each which I'm 90% confident they will win, then I should be feeling much less confident. The same idea applies here with logical consequence. If you believe A1 and A2 to 0.9, then you ought to be disposed to believe the logical consequence of A1 and A2, B, to at least degree 0.8. If B is entailed by A1-A5, each of which you believe to degree 0.9, the overall uncertainty builds up, such that you only ought to believe to believe B to at least degree 0.5.

*N5** will now allow us to deal with cases with multiple premises and account for partial credences. As already mentioned, I don't take this to be the final word on the issues surrounding the normative role of logic, but it gives us a workable principle fit for my purposes here. With a normative role for truth and logic in hand, we can continue to assess the level of constraints a logico-semantic theory places on a cognitive theory.

3. A Supervaluationist Case Study

I will examine the different strengths of connection outlined in the previous section by using supervaluationism as a case study. Supervaluationism can be understood as a family of views which share the same basic machinery. This machinery, roughly speaking, is a set of 'admissible precisifications'. A precisification precisifies language, putting sharp boundaries in the extension of vague words, making a *classical* valuation. On any particular precisification, it is either true or false of everyone that they are 'tall', 'bald', 'young' etc. Many different theories have been developed using this framework, mainly as a model of vagueness. It is possible to create different semantic theories based on this same framework of admissible precisifications. Williams (2014: 393) picks out four different semantics, three of which I'll consider here:

Standard supervaluationism: 'Truth' is truth on all admissible precisifications, 'falsity' is falsity on all admissible precisifications, and indeterminacy is truth on some but all admissible precisifications

Degree supervaluationism: 'Truth' comes in degrees, corresponding to the degree of admissible precisifications a proposition is true on (degree 1 for truth on all admissible precisifications, degree 0.5 for exactly half of the admissible precisifications etc.)

Subvaluationism: 'Truth' is truth on *some* admissible precisification; 'falsity' is falsity on *some* admissible precisification. Propositions true on some but not all admissible precisifications are true and false

These are three well known views, which have been discussed and defended in various places (See e.g. Williams 2011, Keefe 2000, Hyde 1997 for discussion for each theory respectively). They have been differentiated based on what each call 'truth', or 'falsity', and many philosophers would be happy to take them as genuinely different positions. However, Williams worries that if that's all there is to the difference between, we run the risk of losing

any substantive distinction between them. He suggests that to take these theories as genuinely distinct positions, we must also attribute to each a different theory of mind:

> I say above that the four members of the supervaluational family described above are distinct positions. But as I've presented them, they share an underlying semantic machinery, differing mostly on which constructs are picked out as "truth" or "logic." What difference does this make? I think it is exactly that they indicate differences in the cognitive role associated with the various statuses made available by the bare supervaluational framework. (2014: 393)

For Williams, what generates genuine distinction between two theories that utilise the supervaluationist machinery is that the semantics are indicating a different cognitive theory. It is here that it appears that Williams may hold something like my strong interpretation of the commitments non-classical theories hold in the theory of mind: we need to know what theory of mind is attached to a semantics in order to understand the view fully; the theory of mind and the semantic theory are not totally distinct. He suggests the following theories:

Standard supervaluationism ToM: $b(A)= 1$ iff A is supertrue (true on all admissible precisifications), $b(A)= 0$ otherwise

Degree supervaluationism ToM: $b(A)= d$ iff A is true on portion d of admissible precisifications

Subvaluationism ToM: $b(A)= 1$ iff A is sub-true (true on some admissible precisification), $b(A)= 0$ otherwise

These are of course not complete theories of mind, as described in the first section of this chapter. They fill in part of the story about rational belief, identifying what attitudes to take towards each semantic status, a complete theory would also account for decision making, desires etc. Above shows that we can attribute different theories of rational belief to each semantic theory, which delivers different theories of mind for each. Here, it is suggested that standard supervaluationism indicates full belief in supertruths, disbelief in everything else, degree supervaluationism indicates a degree of belief corresponding to the portion of admissible precisifications the relevant proposition is true on, and subvaluationism indicates full belief in anything sub-true, and disbelief otherwise.

Williams suggests that we should associate these cognitive views with each respective semantic theory, and that the semantics may only be verbally distinct if we don't. This is partially a denial of *weak commitment*, the sort of view perhaps seen in Schiffer and Keefe, according to which semantic theories needn't be thought to constrain cognitive theories. But

it also looks like Williams is in line with *strong commitment* here, suggesting that these are the appropriate theories of mind to associate with the semantic theories, and isn't obviously open to the suggestion that there may be others. I endorse the denial of *weak commitment*; for the remainder of the chapter I'll argue that we must think that what makes the semantics different in any interesting way is that they do indicate different cognitive roles, and any difference is immaterial without that. In the following chapters I will argue defend *medium commitment*, and argue that there can be more than one theory of mind associated with particular logico-semantic theories.

3i. Rejecting 'weak commitment'- semantics

I agree that unless we take 'truth', outlined in the three 'different' semantics above, to have a cognitive role, it is not clear that the differences between the three theories are more than verbal. That is to say, the theories disagree about which structures deserve to be *called* truth, but not much more. I suggest that if we take two of the different supervaluational semantics outlined, but blur their theory of mind, we may really only have one substantive theory here, expressed in different ways. This is because truth does come with a cognitive role, as discussed in the previous section, it is the aim of belief. Thus, what is labelled as 'truth' in a semantics should play that role. If it doesn't, then we can examine what does play that cognitive role, i.e. is the aim of belief, and ask which structure really deserves to be called 'truth'.

Let's make this clearer with an example. Consider Rowling, a borderline case of a tall woman. Let's say she is tall on exactly half of the admissible precisifications of 'tall'. On *standard supervaluationism*, Rowling is a borderline case of tallness. 'Rowling is tall' is indeterminate and lacks truth altogether. On *degree supervaluationism*, the same proposition is true to degree 0.5, as it's true on half of the admissible precisifications. There is a legitimate worry here about what difference this really makes, unless it is meant to make a difference to our *cognitive* attitudes. Imagine a case where a degree supervaluationist and a standard supervaluationist claim to have a different semantics, but the same patterns of belief. Say Rowling accepts degree supervaluationist semantics, but thinks that one should have no credence whatsoever in any proposition which is true to a degree less than 1, and full credence in any proposition true to degree 1. This might seem reasonable enough; although truth does come in degrees, we should only invest any belief in fully true propositions. Compare this to Pratchett, who accepts standard supervaluationist semantics, and holds that

you should only have credence in supertrue propositions, and no credence in any proposition which lacks supertruth. Rowling and Pratchett will now have full credence in all of the same propositions, and no credence in all of the same propositions. They will both have no credence in 'Rowling is tall', for example. What difference does it make that Rowling can distinguish between degrees of indeterminacy when Pratchett doesn't? I suggest it makes none. Rowling may be able to *call* a proposition 'true' to a certain degree, while Pratchett cannot, but this seems insignificant if they have the same attitude towards that proposition.

It becomes mysterious what degrees of truth are for Rowling here, if they are disconnected from a cognitive role. Based on her theory of ideal beliefs, she is indistinguishable from Pratchett. Both agree that only supertruths should be assented to, and one should invest zero credence in any proposition that lacks it. What role do degrees of truth play here? What does Rowling gain from being a degree supervaluationist, rather than a standard one? Perhaps one answer might be that we should distinguish degrees of truth as that is how truth is *metaphysically* structured. There is some metaphysical property of degrees of truth that can be assigned to propositions, corresponding to the portion of admissible precisifications they are true on, however only propositions with the property of full truth should be believed.

This would not be a satisfactory answer. Pratchett can also distinguish the property of being true on certain portions of admissible precisifications, but chooses not to call it truth. Any supervaluationist can accept that such properties exist, but the question is, why is it deserving of the name truth? The answer must be that these properties have a *cognitive* role, perhaps like the one suggested by Williams above, that credence in a proposition should match its degree of truth. There is no good answer for the degree supervaluationist of Rowling's sort. I suggest that on her theory there *is* a property deserving of the name truth, but is not the one that comes in degrees here. It is the one playing the appropriate cognitive role, the aim of her belief, *supertruth*. If she chooses to only endorse investing full credence in super truths, then her theory is not substantively distinct from a standard supervaluationist's like Pratchett.

We could make the same point for a subvaluationist. Williams suggests that we should interpret subvaluationism as advocating full credence in any proposition which isn't superfalse, as 'truth' is sub-truth. It might look like this clashes with some other interpretations of subvaluationism, such as Dominic Hyde's (2008: 93-103). Hyde suggests that subvaluationists can take truth (falsity) *simpliciter* to apply to those propositions which are true (false) on some admissible precisification. But take *determinate* truth still to be truth

on all admissible precisifications, and determinate falsity still to be falsity on all admissible precisifications. Indeterminate propositions are those which take on both truth and falsity (unlike in standard supervaluationism, where it is a proposition which takes on neither). With the semantics set out like this, the theory could suggest that we should still only have full credence in propositions which are true on all admissible precisifications, as those are the ones which are *determinately* true. Hyde then might claim to take a theory of mind similar to Pratchett (the one Williams suggests is suitable for the standard supervaluationist), whilst also claiming to be a subvaluationist.

In such a case, Hyde and Pratchett both take full credence in the same propositions, those which are true on all admissible precisifications, and take no credence the remaining propositions. Both have indeterminate propositions, the only difference seems to be that what Pratchett *calls* a truth value gap, Hyde *calls* a truth value glut. Here we perhaps see why Fine says 'This battle of gluts and gaps may be innocuous, purely verbal. For truth on the gap view is simply truth-and-non-falsehood on the glut view and, similarly, falsehood is simply falsehood-and-non-truth (1975, in Keefe and Smith, 1997: 121).

We can see the same question we put to Rowling arising now for Hyde. What do we gain from being a subvaluationist here? What is the importance of calling propositions true on some but not all admissible precisifications true and false? Again, the answer 'this is what the metaphysics of truth are like' would not be satisfactory. Pratchett can also acknowledge that there is a property of being true on some admissible precisifications and not others, but what makes it so that such propositions should be called *true,* and *false*, opposed to neither? Again, the only answer is based in the cognitive role. It should be that if propositions true on some admissible precisifications are to be considered *true,* it is because they are to be *believed*. Why should sub-truth be considered as 'truth' here? I suggest that again in this case, as was the case for the degree supervaluationist, that there *is* a property deserving the name truth, but it's not sub-truth; rather it's the property that is playing the interesting cognitive role and is the aim of belief, supertruth.

We can't straightforwardly take semantic theories to not place any restrictions on related cognitive ones. Truth has to play a cognitive role, and if it doesn't, it becomes unclear what its significance is.

3ii. Rejecting 'weak commitment'- logic

At this stage, someone might object that Hyde and Pratchett *do* endorse different theories, as even though their theories of mind are similar, their semantics lead to different logics. As the real Hyde in fact argues, the difference between sub- and super- valuationism can't be merely verbal, "the validity of modus ponens, for example, depends on it" (2008: 100). Over the bare supervaluationist framework, different notions of consequence are definable. There is the supervaluationist notion; preservation of supertruth (assuming global consequence), and there is the subvaluationist notion; preservation of sub-truth. And isn't each definable without having to say anything about the theory of mind?

At first glance we might think that Hyde and Pratchett could adopt the same theory of mind, whilst Hyde also adopts subvaluationist consequence and Pratchett supervaluationist consequence, and that *this* would result in two importantly distinct theories. I'll argue that this is not a genuine option. As we've discussed, logic also has a cognitive role to play, through the normative role of logic. Yes, subvaluationism and supervaluationism give different notions of consequence, and the latter has modus ponens whilst the former doesn't, but in order to take either notion of consequence seriously, one must take seriously the idea that they are preserving *truth,* which will place cognitive constraints. If Hyde proposes only taking positive credence in supertruths, then what inferences are actually valid for him, and what sets of beliefs are consistent will not be modelled by the subvaluationist notion of consequence, but the *supervaluationist* one. The appropriate set of norms are those placed supervaluational logic, not subvaluational.

Let's define the consequence relations more precisely. For the supervaluationist it is preservation of supertruth:

$\Sigma \vDash_{spv} P =_{df}$ if for all $Q \epsilon \Sigma$, 'Q' is true on all admissible precisifications, then P is true on all admissible precisifications.

On subvaluationism, consequence is defined as preservation of sub-truth:

$\Sigma \vDash_{sbv} P =_{df}$ if for all $Q \epsilon \Sigma$, 'Q' is true on some admissible precisification, then P is true on some admissible precisification.

Let's assume that Hyde wants to have subvaluationist consequence, but to also maintain a theory of mind like that of Pratchett, with positive credence only in supertruths. We see that this is not possible, as the inference patterns that are valid are disconnected from what sentences Hyde actually advocates believing. Consider:

A, A→B $|\neq_{sbv}$ B

A, A→B $\models_{spv}$ B

Recall my proposed characterisation of the normative role of logic, *N5**. We can see that Hyde should take the supervaluationist norms for his beliefs, not the subvaluationist ones. Suppose he believed A, and A→B, and B was brought into question. Should he believe B? According to subvaluationist consequence, the answer is not necessarily, as it is not guaranteed to be true. According to supervaluationism, the answer is yes, as it is guaranteed to be true, and would be irrational if he didn't. Hyde, endorsing the subvaluationist consequence relation, should look at his logic and note that as the truth of B is not guaranteed by the truth of A and A→B, he might make a mistake by inferring it. But, the only instance he would advocate believing A and A→B is if they were *supertrue*, and if they were supertrue, then B would always be supertrue also. So, Hyde could *never* make a mistake by inferring B. As Hyde identifies supertruth as the aim of belief, the consequence relation that preserves supertruth is what delivers the right norms for him, not the relation that preserves sub-truth.

Another way to draw the issue is in terms of consistent belief:

A, A→B, ¬B $|\neq_{sbv}$ Ø

On the subvaluationist notion of consequence it is not inconsistent to accept those three sentences at the same time, but it is on supervaluationist consequence:

A, A→B, ¬B $\models_{spv}$ Ø

So, Hyde's endorsement of subvaluational consequence would suggest that he can accept those three sentences at the same time in some situations. However, his endorsement of supertruth as the aim of belief clashes with this; if Hyde only accepts supertruths, under no circumstances can he simultaneously accept A, A→B and ¬B; Hyde's beliefs are actually constrained by the supervaluationist notion of consequence! The subvaluational consequence relation delivers inappropriate norms for Hyde; it is the supervaluationist relation that correctly models rational belief for him.

This is because in taking supertruth to be the aim of belief, it is supertruth that is playing the truly interesting 'truth' role, rather than sub-truth. Thus, the epistemic norms constraining his

beliefs aren't the ones from a subvaluationist logic, but the supervaluationist logic, which also treat supertruth as 'truth'.

The upshot is that what logico-semantic theories identify as 'truth' and 'logic', comes with at least some constraints on cognitive theories. We have seen that a degree supervaluationist and a subvaluationist aren't free to accept the semantics, and say *anything* about rational belief. Doing so risks stopping their theory from being substantively different from others. Thus, I reject *weak commitment.*

4. Conclusion

I have argued here that accepting a logico-semantic position will come with some constraints on the theory of mind. The idea at the outset was that to take successful steps with the experimental approach, we needed to be able to say what the cognitive commitments of any particular theory of vagueness are, to be able to justify a claim some empirical result is best interpreted in the light of a particular theory of vagueness. It isn't enough just to recognise theories of vagueness as *descriptive*, we needed to be able to say what the significance of the notions they use to model vague language use, such *indeterminacy* or *partial truths*, are in the experimental context. The first step towards this goal was to establish the connection between logico-semantic positions and theories of mind. This step has been taken, *weak commitment* is rejected, thus the two kinds of theories place some kind of constraints on each other, based on the cognitive significance of 'truth' and 'logic'. The remaining question is whether *medium* or *strong* commitment is right. We've seen advocates of particular semantics are *constrained* in what they can say about the theory of mind, but just how constrained are they? Will the semantics be open to more than one theory of mind in some cases or does a logico-semantic position commit you to something very particular in the theory of mind.

In the following two chapters I seek to take the next important step for the experimental approach; having argued that we can take theories of vagueness to have consequences for what they say about belief, we need to determine what these consequences *are* for theories of vagueness. I will examine candidate theories of mind for two mainstream theories of vagueness, supervaluationism and degree theoretic semantics. In the process of this, I will also show that there is more than one theory of mind for each position, and thus that *medium commitment* is the right view.

Chapter 5 Supervaluationism and Suspended Judgement

The last chapter argued that accepting a logico-semantic framework will come with at least some restrictions on an associated theory of mind. The upshot is that theories of vagueness, giving candidate semantics of vague language, are also committed to saying something about the beliefs of competent speakers. In this chapter I will examine what these commitments look like for the supervaluationist, arguing that two options present themselves.

Supervaluationist frameworks give a popular means for dealing with the problems of vagueness, and the features of different supervaluational logics and semantic theories have been well explored. However, the consequences of such frameworks on theories of belief have not been given the same attention. It is well known that a standard supervaluationist theory explains borderline cases as falling in an indeterminate range of a words extension, solves the sorites paradox by claiming that it is false that a very small increase in (say) the number of hairs on a bald man's head would always leave him bald, whilst maintaining that there are no falsifying instances, and holds that it is true of *every* man that they are bald or not bald, even though neither disjunct is true of some men. It is not very well known what all this means when it comes to attitudes in cases affected by vagueness; what *attitude* are we to take towards borderline cases in a supervaluationist setting? The answer I want to put forward here is a *suspension of judgement.* This comes out the most natural attitude given the supervaluational logic and semantics, and especially given the view of the nature of indeterminacy standardly associated with supervaluationism; vagueness as semantic indecision.

At the outset, we need some sort of understanding of suspension of judgement to work with, and I'll suggest that we take it to be represented by a meta-cognitive attitude. With this analysis in hand, we can consider what restrictions *global supervaluational logic* places on attitudes to borderline cases. Both full belief and partial belief are not options, as they cannot be reconciled with a normative role for the logic. Disbelief and suspension of judgement are both compatible, but there are some reasons to prefer suspension. I will go on to argue that taking vagueness to be semantic indecision gives reason to take suspension of judgement as the most appropriate attitude to borderline cases, adding further weight to the argument for

those that accept the traditional link between supervaluationism and that analysis of vagueness.

1. Suspension of Judgement

Before discussing two theories of mind consistent with supervaluationism, I must determine what is meant by suspension of judgement (or just 'suspension' as I may refer to it). It is often included in a traditional tripartite epistemology, but not always afforded serious analysis. I will consider three sorts of accounts here, non-attitudinal, meta-cognitive and Jane Friedman's theory of suspension as inquiry. I defend a meta-cognitive view, but any of the first three are arguably compatible with my aims in this paper. Another set of views about suspension involve taking it to be a sort of partial credence (e.g. Van Frassen 1998, Sturgeon 2010). As will become clear, viewing suspension as partial credence is not compatible with the supervaluationist framework, and thus I leave these views aside here.

When do we suspend judgement? Or withhold belief, or become agnostic? (I assume here that these are just different ways of getting at the same state/attitude.) Standardly we take this stance in cases where we lack sufficient evidence or reason to fall one way or another on some question, when we are unable to appropriately settle some matter or make a judgement. A paradigm case is that of the religious agnostic, who believes there isn't sufficient evidence or argument to either believe or disbelieve that God exists. Unconvinced by millennia of arguments on both sides of the debate, the religious agnostic wants to adopt some sort of neutrality about whether God exists or not, an intentional suspension of believing one way or another. The same can happen with many other questions. We might suspend judgement about historic questions, 'how many steps did I take last year'? Or questions about the future, 'how many steps will I take this year'? I want to suggest here that we could also suspend on questions relating to borderline cases, at least from the supervaluationist perspective; 'Is John bald'? where John is indeterminately bald. These are cases that our epistemic (or some other) circumstances fall short of allowing us to feel comfortable making a judgement.

But what does a suspension of judgement amount to? A common first thought is that we are suspended on some proposition P where we don't believe or disbelieve P, reflected here by Chisholm, for instance: 'H is certain for S at t =df (i) Accepting h is more reasonable for S at t than withholding h (i.e. not accepting h and not accepting ¬h) …' (1976: 27). Call this a non-attitudinal account, for which suspension of judgement is at least partly characterised by

lacking attitudes towards relevant propositions. As Friedman (2013) points out, these sorts of cases tend not to look sufficient for suspension of judgement. Here are some candidate accounts we can reject (as does Friedman):

i. Suspending judgement on P is not just 'P-non-belief'

One suggestion is that subject S suspends judgement about P at time t iff S is in a state of P-non-belief at time t.

'P-non-belief' is having no belief about P or the negation of P. In this case not having a belief about a proposition or its negation is what it is to suspend judgement on that proposition. But that clearly isn't sufficient. For example, say that at time t, I discovered that in 1960 Mikhail Tal competed for the world chess championship. At t-1 a second prior, before I knew the event ever took place, I had no belief about whether or not Mikhail Tal won the 1960 world chess championships. I did not believe Tal won, and I did not believe Tal did not win. However, I could not be said to have been suspending judgement on the matter, I simply had never considered it. It is not plausible that I could be said to be suspending judgement on something which I am not aware of.

ii. Suspending judgement on P is not P-non-belief plus having considered the matter

To avoid the worries encountered there we might suggest that subject S suspends judgement on P at time t iff subject S has considered P and is in a state of P-non-belief at time t.

Again, this doesn't look sufficient. Say upon first hearing about the 1960 world chess championship, I begin to consider if Mikhail Tal was the winner of it. But before I can finish deliberating and come to any decision, at time t my friend phones me. Distracted by the phone call, I stop deliberating. In this case, at time t I have *considered* whether Tal won the championship, have no belief about whether he did or did not, yet could not be said to be suspending on the matter.

Friedman thinks that these sorts of examples reveal the intuition that there is some sort of cognitive act that we associate with suspending judgement, which is vacant in these cases.

iii. Suspending judgement on P is not refraining from belief plus P-non-belief

Friedman also argues that the decision to refrain from/withhold judgement is not the sort of cognitive act that can capture suspension of belief. The suggestion would be that subject S

suspends judgement on P at time t iff subject S is refraining from judgement and is in a state on P-non-belief at P at t.

This cannot be sufficient for suspension of judgement as subjects might reasonably withhold belief for many reasons. If I happen to find chess so boring that I cannot bare to entertain any thoughts about it, when asked whether I thought Tal won the 1960 chess world chess championship, after briefly considering it, I may refrain from any belief about whether he did or didn't. Here I refrain from any belief about P and ¬P, am in a state of P-non-belief, but am not suspending judgement on the matter. Merely refraining from belief is not the cognitive attitude that can capture suspension of belief, as there are many reasons a subject might refrain from belief without suspending judgement on the matter.

iv. Suspending Judgement is not principled non-belief/refraining from belief

To get around the worries related to the previous case, it might be suggested that suspension of belief is refraining from belief for the for the *right* reasons. It's not clear how to cash out what the right reasons would be; for instance, it seems that in the previous case I had some reasons for refraining from beliefs about chess, i.e. that I find it intolerably boring, but these reasons aren't sufficient for suspension of judgement.

Friedman suggests the most promising direction to go is the suggestion that subject S suspends judgement on P at t iff S is in a state of P-non-belief/refraining from belief about P for *epistemic* reasons at t. Interestingly Friedman's concerns with this suggestion are about necessary conditions, unlike the previous cases where she produce counterexamples to show that the suggestions weren't sufficient. Friedman argues that epistemic reasons are not necessary for suspension of belief. We can imagine that at time t1 subject S refrains from belief in whether Martians exist as there is no way to verify whether they do or not. On the current candidate theory, S would be suspending judgement on the matter of the existence of Martians. At t2, Nasa announces they have discovered how to verify whether Martians exist or not, and S loses their epistemic reason for refraining from belief. Friedman thinks that just because S has lost her reasons for refraining from belief at t2, that isn't enough to assume she would have stopped suspending. Even if S can see that her reasons are gone, that doesn't entail that something has changed for her about whether Martians exist or not. If it's right that S was suspending at t1, it seems wrong that she must have stopped suspending by t2 in virtue of having lost her reasons- something about her *attitude* towards Martian life needs to change.

Freidman's focus on necessary conditions are interesting here, as she objects that the previous accounts don't look sufficient; this suggests to the reader the possibility that refraining from belief for epistemic reasons might be sufficient for suspension (this is something she doesn't discuss). Presumably she'd argue that this account is also not sufficient, as it doesn't necessarily include the right kind of attitude towards P, or if it looks sufficient it's because it *does* at least implicitly involve the sort of attitude she has in mind. The issue may be that a compelling counter example to the sufficiency of the above proposal is difficult to construct. Regardless, there is *something* to Friedman's necessity concerns, and an alternate account of suspension can better deal with those problems and can claim to account for why principled non-belief might look sufficient, that it is implicitly involved in such a case. The meta-cognitive account, to be considered next, could play such a role.

Whilst I don't think that non-attitudinal accounts of suspension of judgement are the right way to go, it is worth pointing out at this stage that this is not essential to my overall aims in this paper. If a supervaluationist disagreed with the arguments above and wanted to defend a non-attitudinal version of suspension of judgement, the arguments later to come could still sway them to accept that suspension is the most appropriate attitude to borderline cases. I do maintain that it is best to think of suspension as constituted by an attitude and favour the meta-cognitive account.

1i. The meta-cognitive account

One sort of account is what we'll call the *meta-cognitive* account. The proposal is that a suspension of judgement is constituted by an attitude towards our own attitudes; that we suspend judgement on P when we believe that we cannot make a judgement on P/neither believe nor disbelieve P. The suggestion I have in mind can be seen in the following passages:

> Withholding p, then, is a propositional attitude distinct from mere failure to take up any attitude towards p. Like believing or disbelieving, it is taking an attitude towards a proposition. What more can one say about withholding? As I shall be using the term, withholding p involves resistance, voluntary or involuntary, to believing p and to disbelieving p. The only thing one must consider in order to believe p or to disbelieve p is p (or its denial). But to withhold p (in the sense I have in mind) one must, in addition, consider the prospect of one's believing p as well as the prospect of one's disbelieving p; otherwise one will not be able to resist both believing p and disbelieving p. So withholding p involves not only an attitude towards p but also attitudes towards attitudes towards p. (Bergman 2005: 421)

> The suspension of belief, or the reservation of judgement, has a degree of cognitive sophistication beyond mere believing, in that it arguably involves the ability to have beliefs

> about what one does not believe. This is because a deliberative process that aims to establish whether or not p, must terminate, as long as it is not interrupted, in the formation of some judgement or belief about the matter in question, about whether p. When the deliberative process terminates in a reservation of judgement with respect to p, the belief consequently formed about the matter is the belief that one does not know whether or not p. Suspension of judgement necessarily involves thoughts about one's own epistemic perspective on whether or not p, namely, that one's epistemic perspective falls short of establishing whether p and thus that one does not know whether p. Now, obviously, in this kind of case, to believe that one does not know p is not merely to think that one does not have enough evidence strictly speaking to claim to know p but that for all that one still believes p. Rather, in this kind of case, to believe that one does not know whether p is to believe that one neither believes p nor disbelieves p. (Crawford 2004: 226)

The proposal is that suspension of judgement requires not only to not believe P or disbelieve P, but also a meta-cognitive attitude, a belief that one does not believe or disbelieve P. Put more precisely:

Meta-cog: Subject S suspends on P at time t iff subject S neither believes nor disbelieves P at time t, and simultaneously believes that she neither believes nor disbelieves P.

Bergman and Crawford don't give precise formulations of the proposal, but I take it that *Meta-cog* gives a fair interpretation of what they have in mind, particularly when Crawford says: 'Reference to oneself and one's lack of beliefs is not merely a natural but optional way of describing what is going on; it is an essential feature of what it is to suspend judgement. In order to suspend judgement about whether p, it is, I am suggesting, necessary to believe that you do not believe or disbelieve p'. (pp. 226)

This isn't necessarily the only candidate for this type of theory. What is most important is that there is an essential meta-cognitive component to suspension of judgement, the reference to oneself and one's lack of beliefs as Crawford says above. This attitude should represent the indecision about the relevant propositions, but it needn't be that believing you don't believe nor disbelieve P is the essential meta-cognitive attitude. I think just merely recognising that you don't believe or disbelieve P may miss something important about suspension, that it usually is a case where you believe you *shouldn't* believe or disbelieve P, or that you are not in a position to make a judgement either way. We might formulate a second account this way:

Meta-cog:* Subject S suspends judgement on P at time t iff S neither believes nor disbelieves P, and believes that she cannot make a judgement on P, at time t.

On this sort of account, it will likely still be the case that subject S believes she doesn't believe or disbelieve P when she suspends judgement on it (unless it's a very strange case where they have inconsistent beliefs and are not aware of it), it's just that the essential meta-

cognitive attitude for suspension is not that one, but the belief that one cannot make a judgement about P. This is the direction that looks right to me, but there are more alternatives. One might suggest an account on which both above attitudes are necessary for example, or there could be separate meta-cognitive attitudes that someone thinks do a better job. We can think of a generalised meta-cognitive theory, on which suspension of judgement is neither believing or disbelieving P plus having the appropriate meta-cognitive attitude. Again, any of these would be fine for my overall aims in this chapter, but I'll work with *meta-cog**.

One question we might have is whether this is really a *non-attitudinal* account of suspension of judgement; after all, it leaves us in a state of P-non-belief when we suspend on P. The reply would be that whilst we lack first order beliefs about P when we suspend on it, suspension is still constituted by an indecision representing attitude; the meta-cognitive attitude. This is what is essential to the account and gives a way to claim that it deserves to be considered an attitudinal account. I think that is probably right, but again nothing in my argument necessarily turns on that. If someone thinks this really is a non-attitudinal view, then it still looks like an interesting and viable account of suspension of judgement nonetheless.

Let's see the suggestions in practice. If I'm asked whether it will rain tomorrow, I may deliberate over it, but ultimately decide that I don't have enough evidence to believe that it will or won't and suspend judgement on the matter. In this case I neither believe nor disbelieve that it will rain, and at the end of my deliberative process, I have also formed the meta-cognitive belief that I cannot make a judgement on whether it will rain or not. In the case of vagueness, a subject suspends on 'John is bald' iff she neither believes nor disbelieves that John is bald and *believes* that she cannot make a judgement on it.

This seems to account for what's going on with the paradigm case of the religious agnostic. After deliberation over the matter and considering the evidence, the agnostic has not made a first order judgement about whether god exists, but a meta-cognitive one, taking up the belief that she shouldn't make a judgement on the matter.

It also handles the problems that non-attitudinal accounts had, showing us what went wrong with those cases. It was suggested having considered a matter and P-non-belief is not sufficient suspending on P, because a subject might begin deliberating and be interrupted; the meta-cognitive account can expand on this, claiming it is because the subject did not take up

the appropriate meta-cognitive attitude to P. Friedman suggested the subject that 'suspends' over the existence of Martians for epistemic reasons can lose her reasons without necessarily stopping suspending judgement, the meta-cognitive account can fill that picture in: this is because losing her reasons doesn't necessarily mean she has changed the meta-cognitive attitude representing her suspension. It can also explain why such an account might *look* sufficient for suspension of judgement; deciding to have no belief in P for epistemic reasons is very similar to the meta-cognitive account, it looks quite likely that in those cases you will have something like the belief that you shouldn't make a judgement on P.

I don't see simple ways to object to the meta-cognitive account. Is it sufficient for suspension of judgement? If it is not, then someone should be able to point out what is missing from the analysis of religious agnostic; it is not clear to me what that would be. It is necessary? Well the straightforward removal of the meta-cognitive attitudes would just leave us with a non-belief account, which we have seen is problematic. An argument that it is not necessary would require supplementation with an alternative analysis of suspension of judgement. I will examine and reject two more here and suggest that the meta-cognitive account gives a successful and common-sense view of suspension of judgement.

1ii. Suspension as Inquiry

Jane Friedman (2015) presents her own account of suspended judgement, which ties it closely to *inquiry.* She doesn't offer a particularly formal version of the proposal, but she is arguing for something like the following biconditional: subject S is suspending judgement about question Q iff S is inquiring into Q.

Both sides of this biconditional need elaboration. Friedman argues that suspension of judgement falls in a class of *interrogative attitudes* (IAs). Typical IAs include wondering, deliberating, contemplating, investigating and curiosity. These are question directed attitudes; we 'wonder' or are 'curious' about questions, wondering about whether it will rain tomorrow. The suggestion is that suspension falls into this category, hence talking about suspending about a question Q above; so, on this account suspension of judgement is not a propositional attitude.

To 'inquire' into question Q on this account is to have Q on your *research agenda,* which 'record our epistemic goals by way of the questions we wish to answer' (2015: 309). We are

inquiring into Q when Q is a question we want to answer, and we are at least minimally sensitive to information that bears on the question.

The general suggestion is that a subject suspends on a question wherever she puts that question on her research agenda, i.e. has an IA towards it (which may just be suspension), aims to answer it, is at least minimally sensitive towards information that bears on it. I'll consider arguments for each direction of the biconditional in turn, and it will clarify more of how the proposal is meant to work.

First the right to left, that is, that the claim that if a subject is inquiring into a question, then they are suspended about it. Friedman's argument here relies on the tension between having an IA towards a question, and knowing or believing an answer to it. For instance, there seems something inappropriate about both *knowing* who won the 1960 world chess championship, and also being *curious* about who won it. Any instance of knowing some proposition P and also wondering whether/investigating/being curious whether P seems infelicitous. Friedman suggests that there is a norm against both knowing whether Q, and having an IA towards Q. She suggests that this is because there is a conflict of attitudes here, and asks what the source of it could be. Her answer is that knowledge entails belief, IAs must entail an attitude which conflicts with belief: *suspension* of belief. She claims this is the best way of explaining why it is inappropriate to inquire into a question you already know the answer to.

Freidman is more cautious about the left to right direction of the biconditional; the claim that if a subject is suspending then they are inquiring; suspending judgement on a question requires putting that question on their research agenda. One point she makes is that there are cases where suspending on a question looks inappropriate exactly where further inquiry does, claiming if we can't put a question on our research agenda, then we can't suspend on it. For instance, we can't inquire into 'what colour was Thomas Jefferson's Ferrari?', as it contains a false presupposition. Similarly, it looks inappropriate to suspend on this question, as we know Thomas Jefferson didn't own a Ferrari. One way of explaining why we can't suspend on such a question is that suspending requires putting a question on your research agenda.

Another point she makes in favour of this direction of the conditional is that it lines up with sceptical and pragmatist notions. A sceptic, not satisfied they're justified in belief, has suspended judgement, and accordingly is inquiring. Sextus Empiricus, for example, claims other philosophical schools believe they have discovered the truth, but that the 'sceptics are still investigating' (2000: 1.13). Similarly, the Peircean notion of doubt is related. The idea

being that that when are in doubt about something we are pushed to inquire until we can resolve it. Friedman wants to claim that suspension of judgement also comes with this goal-oriented push to inquiry and resolution. Friedman suggests that all IAs come with this goal-orientation, i.e. wondering, investigation, deliberating, and as suspension of judgement is 'entailed' by all IAs, we might think that suspension of judgement is a *core* IA. 'Qua core IA, it isn't a stretch to think suspension shares this property' (2015: 318).

I think it is plausible that suspension of judgement and inquiry are closely related. I took it from the outset that suspension comes when a subject is unable to settle some matter; some sort of epistemic openness to such a matter seems appropriate. However, as it stands Friedman's analysis doesn't look quite right.

The features of suspension of judgement are still somewhat unclear on this view. Friedman says that suspension is a 'neutral *contentful* attitude' (pp. 322). But it's not clear what the content of this attitude exactly is. For example, she also indicates that suspension of judgement is meant to be a doxastic attitude, of the sort included in a traditional epistemology (pp. 330). It isn't clear how these aspects of suspension are meant to be reconciled with its role *qua* interrogative attitude; it isn't a propositional attitude, but a question-directed one.

This concern comes to light when criticising some of Friedman's arguments. For instance, I think the argument for the left to right reading of the biconditional comes to something of a non-sequitur, perhaps at least without some more clarification. The suggestion is that it is inappropriate to know the answer to question and to have an IA towards it. Friedman says we can explain this by suggesting that knowing implies belief and inquiry entails suspension of judgement. But if suspension of judgement is meant to be a sort of IA too, or even the *core* IA, then why does it help to explain this tension? Why does it explain the tension when other IAs cannot? Presumably because of something to do with its *doxastic* role, but it's not clear what features it has as a doxastic attitude, if any. Presumably it is meant to be about an epistemic openness to the question or relevant propositions, but I don't think it's enough to just be told that this is a state that implies openness and is 'distinct from merely lacking beliefs in some answers'. The doxastic features remain perplexing at this stage, what seems most sensible is to reinforce the account with a theory that *can* explain the doxastic features; meta-cognitive account could do just that. A more promising way to characterise the above tensions may be to say that there is a tension between knowing P and inquiring in P, because

knowing implies believe P, and inquiring implies suspension on P, i.e. the belief that you *don't* believe or disbelieve P (or shouldn't make a judgement on P).

Although, it's not obvious this tension ever needed explaining. There could be lots of reasons why knowing who won the 1960 world chess championship and *wondering* who won it is infelicitous, for instance perhaps because it's wrong to inquire when the goal of inquiry has already been achieved. It can just be straightforwardly inappropriate to inquire into a question you know the answer to, the involvement of suspension of judgement isn't required to explain it. It doesn't look like the conclusion that suspension of judgement must be involved in inquiring necessarily follows from the tension appealed to.

Beyond this, there are sufficiency concerns for the account. It was suggested above that the subject that begins to consider a question but is interrupted prior to being able to finish deliberations couldn't have sufficiently suspended judgement, something Jane Freidman endorsed. It looks like this account of suspension of judgement is open to the very same concern; surely the subject who has begun deliberating but has been interrupted *has* put the question on her research agenda. Why is it that this account fairs better than the non-attitudinal account? Friedman does suggest an answer, again that appeals to the role of suspension as a doxastic state, distinct from being mere non-belief. But here we have the same problem of course; what are the features of this doxastic state? Under what circumstances do we have it, if this account can explain it but not the non-attitudinal account? Again, it looks like a simple solution to take the meta-cognitive way out: we have it where we have the appropriate meta-cognitive attitude.

The issue is that it looked right before that the interrupted deliberator hadn't sufficiently suspended yet. This is related to another concern, that we might think that suspension of judgement comes at the *end* of deliberation. For instance, the religious agnostic is a subject that has considered the relevant evidence and arguments, completed deliberations, and decided to suspend on whether god exists. It seems like the agnostic is in a different doxastic state when deliberating and when having decided to suspend. Of course, in both cases she has no belief in the propositions at hand, or is refraining from belief, but as discussed, neither of those are sufficient to capture the idea of suspension we wanted. Friedman's account cannot explain the difference in the two cases; whilst both deliberating and having concluded she must suspend, the religious agnostic has the question of whether god on her research agenda, meaning she is in fact suspending judgement in both cases. This just doesn't look right, there

is a difference between the state of the agnostic in these cases. Suspension as inquiry cannot account for why. The meta-cognitive account is also better on this score; we can point out that the agnostic doesn't have the appropriate meta-cognitive attitude *until* having decided to suspend judgement.

Whilst it does seem right that inquiring and being suspended are closely related, giving an analysis of suspension in terms of being an inquirer is not correct. The meta-cognitive account clearly handles issues it faces much better. As with the non-attitudinal accounts, I do still take it that this analysis will be consistent with my broader aims here. If we are suspending judgement iff we have a question on our research agenda, we suspend on borderline cases when we put the question of whether or not they obtain on our research agenda.

2. Supervaluationism and Belief

With the meta-cognitive analysis of suspension of judgement in hand, we can return to the central question. Consider a standard supervaluationist theory; truth is truth on all admissible precisifications (supertruth), falsity is falsity on all admissible precisifications (superfalsity), indeterminacy is truth on some admissible precisifications and not others, validity is preservation of supertruth. Taking this picture to be the model of competent vague language use, what are the appropriate doxastic attitudes for these semantic statuses, what is the supervaluationist committed to saying about a competent speaker's belief? (Again, I'll still call this the 'rational attitude', but note that it is still a descriptive notion). There are three semantic status to account for; truth, falsity and indeterminacy. It is clear we can associate belief with truth, disbelief with falsity, but it is less clear what should be said about indeterminacy. Is any attitude towards indeterminacy consistent with the supervaluationist framework and semantics, or are there restrictions?

At the outset, I take it that there are four possible options; belief, partial belief, disbelief and suspension of judgement, and I will be defending the view that the most appropriate attitude is suspension of judgement[6]. Although the subject is not always given significant attention, some have considered it. Rosanna Keefe suggests that withholding belief (i.e. suspending)

[6] I appreciate that there is an important question about what the object of belief is on a supervaluationist theory, and what their story about propositions is. Unfortunately there isn't space here for a satisfactory treatment, but I take it that the answer to that question is neutral to the success of my argument. I am discussing what supervaluationism is committed to saying about belief, whatever the object of that belief turns out to be.

and disbelieving both look plausible ways to respond to borderline cases, but suggests that we have in mind *'no clear-cut rule'* about how we *should* respond (2000: 154-5). As seen in the last chapter, Robbie Williams (2014) suggests that disbelief is the attitude supervaluationists should endorse towards borderline cases, but he doesn't consider suspension of judgement, and the viability of that approach. Elsewhere, the possibility is mentioned, for example by Achille Varzi, who writes that the supervaluationist must suspend judgement over borderline cases (2001: 146).

In the following section I'll first argue that belief and partial belief aren't options for the rational attitude to take towards indeterminacy based on this framework. I'll then suggest that both disbelief and suspension remain consistent, but I'll argue that suspension of judgement is the natural choice for the supervaluationist; it comes out the best option based on the bare logics and semantics of the framework, and based the philosophical story associated with supervaluationism, vagueness as semantic indecision.

2i. The logic and semantics

As mentioned, four basic candidates for attitudes towards borderline cases and indeterminacy are belief, partial belief, disbelief and suspension of judgement. However, using the normative role of logic laid out in the previous chapter, belief and partial belief can be ruled out based on the supervaluationist logic.

N5:* Where A*1*, …, A*n* ⊨ B, you ought to see to it that b(B) is greater than or equal to b(A*1*) +…+ b(A*n*) – (*n*-1)

In general, the idea is that credence should be preserved over logical entailment. Given this norm, it follows that supervaluationists cannot have positive credence in borderline cases. First note that global supervaluational logic validates the entailment from P to DP: P ⊨ DP. Williams (2011: 135) argues that given a normative role of logic, our credence in P should not exceed our credence in DP. Where P is indeterminate, we should have zero credence in DP; if some hue is borderline red, it is superfalse that it is *definitely* red, and we can thus disbelieve it. Accordingly, we shouldn't have any positive credence in indeterminate P, as it entails DP, which we disbelieve.

Williams only considers the entailment for P to definitely P, but presumably other examples can be found in which positive credence in borderline cases would violate *N5**. For instance,

assume that P is indeterminate, and Q is superfalse. In this case the following entailment has indeterminate premises and a superfalse conclusion:

$P, P \rightarrow Q \vDash Q$

Again, it is appropriate to disbelieve the conclusion, as it is superfalse. But this means, based on *N5**, it is inappropriate to have any positive credence in both of the two premises. Having positive credence in indeterminate cases isn't consistent with a normative role for supervaluationist logic. This means it is inappropriate for the supervaluationist to endorse either full belief or partial belief in borderline cases, as both involve positive credence. The entailments above (as well as others that can be found) show that only an attitude that involves no credence in indeterminacy is appropriate. So far, that leaves the door open to both disbelief and a suspension of judgement.

There is a further argument that can be made to suggest that disbelief is the only attitude one can take towards borderline cases consistent with a normative role for supervaluationist logic. The suggestion is that just as belief is preserved forward over entailment, disbelief should be preserved backwards. The idea can be captured with something like the following:

N6: Where $P \vDash Q$, if we disbelieve Q, then we should disbelieve P

The argument then would be that, where P is borderline, DP is superfalse. Thus, the only appropriate attitude to take towards DP is disbelief when P is borderline, and as P entails DP, we should also disbelieve P.

This argument strikes me as being just a little too strong, and it can be resisted for this reason. I do grant that there is something plausible about it. Consider a (classical) case with modus ponens:

$A, A \rightarrow B \vDash B$

We might think given this entailment, if we were to *disbelieve* B, we should then disbelieve the conjunction $(A \ \& \ (A \rightarrow B))$. If we don't believe B, then we shouldn't believe a set of premises that would logically commit us to B. Whilst something does seem right about this, I worry the intuition relies on *contraposition*. From the above entailment, we can classically move to:

$\neg B \vDash \neg(A \ \& \ (A \rightarrow B))$

Backwards preservation of disbelief can classically be rephrased as a forward preservation of belief in corresponding negated sentences. But in the case of supervaluationism this doesn't hold, as contraposition is not valid:

$P \vDash DP$

$\neg DP \not\vDash \neg P$

If what is intuitive about *N6* relies on contraposition, then it isn't an appropriate norm in the supervaluationist setting, at least where the 'definitely' operator is in the language, as the logic ceases to be classical. Whilst P does entail DP, ¬DP doesn't entail ¬P, and accordingly it's not so clear to me we should accept as a strong a norm as *N6*, and accept disbelieving DP should commit us to disbelieving P.

A weaker, and perhaps more palatable, interpretation of the norm is just to say that where you disbelieve a proposition, you shouldn't believe a set of propositions that entail it:

N7: Where $P \vDash Q$, if you disbelieve Q, you shouldn't believe P.

This essentially places the same norm as *N5**, our credence in P shouldn't be greater than our credence in Q. Where P entails DP, and P is borderline, we disbelieve DP, so we shouldn't believe P. But this is consistent with both disbelieving P and suspending judgement.

We can still have *N6* in some cases for supervaluationist logic; for 'definitely' operator free sections of the language, allowing us to use the norm in many instances where it does seem plausible. It is still safe to move from disbelieving B to disbelieving $(A \,\&\, (A \rightarrow B))$ for example, as this is a case where contraposition is valid. This is an overall pleasant result, where the language is restricted and the logic is classical we can have *N6*, which does look plausible in many cases. But in the overall non-classical logic, we only have the weaker norm *N7*, which is consistent with both disbelieving and suspending judgement on borderline cases.

So, the supervaluationist framework rules out partial and full belief as options for the rational attitude towards indeterminacy, however both disbelief and suspension of judgement remain consistent options for the framework. I take it that this shows that there is a medium level of commitment between logico-semantic theories and theories of mind: the supervaluationist framework does rule out some theories of mind (any that involve positive belief in indeterminacy), however there remains more than one consistent option. That having been said, I think that there are philosophical reasons to prefer as suspension model based both on

the framework, and the philosophical story standardly associated with it, for the remainder of this section I'll discuss reasons of the former sort.

One reason is that taking disbelief as the attitude to indeterminacy doesn't seem to do our initial intuitions about indeterminacy justice, given both what those intuitions are and that we have them at all. Theorists typically set out borderline cases as those which speakers do not know how to classify, that they don't want to say falls into the positive or negative range of a vague predicate's extension. It is a case that speakers are uncomfortable making any judgement about. This sort of intuition about borderline cases, to which supervaluationist appeal in introducing truth value gaps in the first place, can be better captured by suspending judgement on indeterminate cases than by disbelief. Furthermore, it seems significant that we have intuitions about borderline cases being weird at all. If we disbelieve indeterminate P, then we take the same attitude to it as we do to superfalse cases. But do we really want to suggest we take the same attitude to 'S is bald', when S is definitely not bald, as we do when S is indeterminately bald? This doesn't fit with our intuitions that the two cases *are* different in an important way, but a suspension of judgement on the borderline cases would.

Another thought is based on some of the non-classical semantic results. Supervaluationism validates all instances of the law of excluded middle, even for borderline cases. Accordingly, 'S is bald or not bald' is also supertrue, and it is thus appropriate to believe it, even though in some cases it may not be true of S that they are bald, nor true that they are not bald. Given that we must always endorse 'S is bald or not bald', it looks better to suspend judgement on borderline cases, rather than disbelieve them. Where S is a borderline case of baldness, it looks wrong to accept that S is bald or not bald, but disbelieve that S is bald, and that S is not bald. It is much more palatable to accept that S is bald or not bald, and then suspend over which; suspend judgement on 'S is bald', and 'S is not bald'.

A similar case is the truth of the existence of a 'sharp boundary' of sorts. It comes out supertrue that there is a number of hairs such that someone with that number is not bald, but that the removal of one hair would turn them bald. Accordingly, it is appropriate to believe that such a number of hairs exists. It is of course not true of any particular number of hairs; however, it is indeterminate for some numbers of hairs. Given that it is appropriate to endorse that there is such a number, it looks strange to disbelieve that every candidate could be it. It is better to be *suspended*; to believe that there is such a number, but suspend judgement over which of a class of candidates is that number.

Of course, if we were to accept suspension of judgement as the appropriate attitude towards borderline cases based on the logic and semantics alone, disbelief would still have a related and important role to play in the supervaluationist picture: we must disbelieve that 'S is *definitely* bald' where S is indeterminately bald. Any intuition that we should disbelieve 'S is bald' can be explained as really wanting to disbelieve the former sentence.

So, the logic leaves both disbelief and suspension of judgement as options, but there are reasons to prefer suspension of judgement; it fits better with the intuition that indeterminate and superfalse cases are different in an important way, and fits better with supervaluationist non-classical semantics. I think there is further reason to take suspension of judgement as the most appropriate supervaluationist attitude to borderline cases when considering the theory of indeterminacy standardly associated with it: vagueness as semantic indecision. A supervaluationist may choose not to take this traditional connection seriously, and endorse the logic and semantics without endorsing vagueness as semantic indecision. They would still have reasons to prefer suspension of judgement to disbelief based on my arguments thus far. But many do seem to endorse the connection, particularly David Lewis and Kit Fine, but also Dummett, Keefe and Varzi. It seems to me that those who have even more reason to accept that suspension of judgement is the most appropriate attitude towards indeterminacy.

2ii. Vagueness as semantic indecision

'Vagueness as semantic indecision' takes vagueness to be a purely semantic matter. The world is precise, but our way of verbally representing and conceptualising the world falls short of this precision, following the sort of view put forward by Russell: 'Vagueness and precision alike are characteristics that can only belong to a *representation*, of which language is an example' (1923, in Keefe and Smith 1997: 62). Words vaguely represent precise things in the world, rather than precisely representing vague things.

In this way, we can think of vagueness as a 'deficiency of meaning', for example as Kit Fine did (1975: 265). Vague words have ranges where they clearly apply, and clearly don't apply, but a rough indeterminate range between them. This is modelled with a set of admissible precisifications, a plurality of equally good candidates for a sharp boundary between a range of positive and negative application for vague language. The idea is, as the world is precise, each admissible precisification picks out a different possible precise extension/property/object which could be referred to by a vague word. For instance, the

admissible precisifications for ‘Toronto’ each pick out an equally good candidate geographical locations for the referent of that word. The admissible precisifications of ‘red’ pick different precise candidates for the property carved out by that term. This is why supervaluationism is sometimes spoken about as picturing vagueness as a sort of semantic indecision or ambiguity; vague language leaves it open which precise property (/extension etc.) it is meant to pick out, it remains undetermined between a range of equally good candidates. This is what Fine has in mind when he suggests ‘Vagueness is ambiguity on a grand systematic scale’ (1975: 282), and Lewis when he suggests that vagueness is semantic indecision (1986: 282).

This gives justification for the semantics, propositions which are true on all given admissible precisifications are supertrue, and those false on all admissible precisifications are superfalse. These propositions are not affected by the range of different admissible precisifications; each yields the same semantic value. To borrow an example from Mehlberg (1956), the vagueness of ‘Toronto’ (for the supervaluationist) means our language leaves it undecided which of a range of precise geographical locations is associated with it. Despite this vagueness, it clearly true that Toronto is in Canada, as on every way of carving ‘Toronto’ out, it is in Canada. It is clearly false that Toronto is in Europe, as every way of carving ‘Toronto’ out leave it *not* in Europe. However, some propositions are affected by the range of admissible precisifications. Whether there is an even number of trees in Toronto depends on which geographical area we associate with the term; under some interpretations of ‘Toronto’ the number of trees within the boundaries of the city is even, under the rest it is odd. Thus, we deem the proposition ‘the number of trees in Toronto is even’ indeterminate. Due to the vagueness of ‘Toronto’, and our semantic indecision, we cannot determine the truth value of that proposition. There is more than one precise geographical location that can be associated with that word, and for matters which turn on which one in particular ‘Toronto’ refers to, such as whether the city contains an even number of trees, we cannot determine the answer. If we were ever to precisify ‘Toronto’, and choose one of the admissible precisifications to be *the* precise boundary for the city, then I would be able to say whether the number of trees there is even. But as it stands the matter goes unsettled.

This philosophical backdrop is meant to justify the unusual non-classical semantics generated by supervaluationism that have been discussed. Supervaluationism allows for true disjunctions which lack a true disjunct. For instance, the theory validates every instance of the law of excluded middle: it is true of every man that they are bald *or* not bald, even though

neither individually is true of borderline bald men. Supervaluationism also satisfies some existential generalisations that lack verifying instances. In particular, it is always true that there *is* a sharp boundary in the extension of vague words, however it is never true of any candidate that it is *the* boundary. It is true that there is a number of hairs such that a man with that number is bald, but an extra hair would make him not bald, however it isn't true of any particular number of hairs that it satisfies that description.

These results can look strange or unattractive, but if vagueness is a matter of semantic indecision, then it can be argued that they're the correct results. Consider the following quotes:

> Suppose I press my hand against my eyes and 'see stars'. Then LEM should hold for the sentence S = 'I see many stars', if it is taken as a vague description of a precise experience. (Fine 1975: 285)

> The statement 'That is red' will… be neither definitely true nor definitely false: but, since the object is on the borderline between orange and red- there is no other colour which is a candidate for being the colour of the object- the disjunctive statement, 'That is either orange or red' will be definitely true, even though neither of its disjuncts is. (Dummett 1975, in Keefe and Smith 1997: 106)

The idea is that if vague words refer to precise extensions/sets/properties/etc. in the world, but it's just indeterminate *which* is referred to from a range of equally good candidates, then the results are correct. As 'bald' picks out some precise property, it *is* true that all men are bald or not bald. However, as no property is determinately picked out in *particular*, we are left with borderline cases, men for whom we cannot settle the matter of whether they are bald or not. Similarly, as 'bald' picks out a precise property, it *is* true that there is a sharp boundary to its extension. But as it is indeterminate which property 'bald' picks out, we cannot say which of a range of equally good sharp boundaries this is. Again, in Dummett's example, 'red' and 'orange' pick out precise properties in the world, but it is indeterminate which property that is. It is true that borderline red/orange patches *are* red or orange; as Dummett says, there is nothing else they could be. We just cannot settle the matter of which colour the borderline patches are due to the semantic indecision about the referents of 'red' and 'orange'.

I submit that if the supervaluationist buys into this picture, and accepts that semantic indecision justifies these semantic results, then they should accept suspension as the most appropriate attitude to borderline cases also. If we are compelled to endorse that it is true of every man that they are bald or not, because it is indeterminate which property is picked out

by ‘bald’, then we should also be compelled to suspend on borderline cases: whether a borderline case is bald or not simply depends on the referent of ‘bald’, and as this referent is indeterminate, we cannot settle the matter. The same thought is seen in Varzi: ‘…we must suspend judgment when it comes to saying whether Everest is mostly in Tibet: the truth value of such a statement depends crucially on how much land one includes in the referent of ‘Everest’’ (2001: 146).

The suggestion is that where the semantic indecision of vague language leaves it indeterminate of some object whether some predicate applies, it is most appropriate to suspend judgement on the case. Vague language does refer to precise properties, but it’s not determinate which, and where the indeterminacy becomes an issue, it’s not appropriate to make a judgement on the matter (and we of course also get the intuition that we shouldn’t or can’t). It looks more natural to suspend judgement than to disbelieve given the semantic indecision view; we shouldn’t make the judgement that we disbelieve that borderline Bob is bald. That’s the judgement we make when it’s determinate it that he’s not bald, when comes out bald on all the properties ‘bald’ indeterminately refers to. In the indeterminate case, we cannot settle the matter of whether Bob is bald or not; the deficiency in the meaning of ‘bald’ leaves it undecidable.

Of course, the same arguments from the previous section apply still; suspension fits more naturally with accepting LEM and existential generalisations that lack verifying instances and aligns better with our intuitions about borderline cases. I suggest now that suspension also firsts better with the traditional view of vagueness associated with supervaluationism. We can make this additional point more convincing by exploiting the similarities between ambiguity and vagueness conceived as semantic indecision.

2iii. The argument from analogy with ambiguity

As mentioned above, supervaluationism is sometimes seen as likening vagueness to ambiguity, and as already noted, Fine was very explicit in thinking the two were similar in interesting ways. There are cases where it isn’t possible to determine the truth value of a sentence due to an ambiguity, in which it seems appropriate to suspend judgement. Importantly here, there is an analogy between the ambiguous cases that compel one to suspend judgement and borderline cases, compelling one to suspend judgement on the latter also.

It's well known that 'bank' is ambiguous. There are two distinct meanings associated with the word: a financial institution and a riverside. There is something of a parallel here with the supervaluationist view of vagueness; according to the theory 'bald' has more than one precise extension associated with it. There is also important disanalogy. Keefe suggests that ambiguous expressions have more than one actual meaning, whereas the vague words (typically) have a univocal meaning (2000: 157). This is true. Fine thought about this in the following way:

> Ambiguity is like the super-imposition of several pictures, vagueness like an unfinished picture, with marginal notes for completion. One can say that a super-imposed picture is realistic if each of its disentanglements are; and one can say that an unfinished picture is realistic if each of its completions are. (1975: 283)

The meaning of vague words is underdetermined, as there is more than one extension indeterminately associated with them. Ambiguous words (sentences) are overdetermined, there are many distinct meanings which are each determinately associated with them. But this difference isn't important here. What is important is that sometimes we must suspend judgement on ambiguous sentences, and the similarities between these and borderline cases from the supervaluationist perspective may give us reason to suspend over those also.

Suppose you were completing a (somewhat strange) questionnaire. The first question asks, 'have you been to the bank today?'. You have three boxes you can tick in response: 'yes', 'no' and 'can't tell'. Which box do you tick? Presumably you tick 'yes' if you believe it's true you have, 'no' if you believe it's false you have, and 'can't tell' under some other circumstances[7]. So, what attitude do you take to the sentence 'I have been to the bank today'? It may vary depending on how you disambiguate 'bank'. In cases where we unable to disambiguate the sentence, as we are with the questionnaire (assuming we cannot contact its designer to discover what she meant by 'bank'), we can still make judgements in a similar way to the supervaluationist. If you have been to the financial institution and the riverside today, you can happily endorse 'I have been to the bank today', and tick the box labelled 'yes', as however 'bank' is disambiguated, you know the sentence would be true. If you have neither been to the financial institution nor the riverside, then you can happily disbelieve 'I have been to the bank today' and tick the 'no' box, as it is false on both possible

[7] I do not mean to suggest that ticking 'can't tell' should be exclusively associated with a suspension of judgement. There are many reasons one might select this option in a questionnaire, if they genuinely don't know the answer for instance. I only claim that one of the reasons they would is if they were suspending. In this way, a subject ticking 'can't tell' is not sufficient to say they have suspended, but it is necessary; if they suspend judgement, then they will tick the 'can't tell' option. This conditional is all I need for this section.

interpretations of ‘bank’. What about the case where you have visited the riverside, but not the financial institution, what attitude should you take to the ambiguous sentence in question? The rational thing to do here may be to suspend judgement over it. On one interpretation of ‘bank’ the sentence is false, but on the other it’s true, and you cannot make a judgement until you know which disambiguation is relevant here. Which way should you respond in the questionnaire? You surely shouldn’t tick the ‘yes’ box, as there is an interpretation of ‘bank’ on which it’s false you went. You surely shouldn’t tick the ‘no’ box, as there is an interpretation of ‘bank’ on which it is true you went. It is only appropriate to suspend judgement and tick the ‘can’t tell’ box; you cannot determine how to respond to the question until you know how ‘bank’ is to be disambiguated.

If we find it convincing that in this case we should suspend over ‘I have been to the bank’, due to the plurality of interpretations being problematic, and not allowing us to settle its truth value, there is an argument that we should suspend over borderline cases also; here we also fail to settle truth values due to the plurality of interpretations. Imagine that having ticked ‘can’t tell’ on the questionnaire for the first question, you proceed to question two: ‘are you bald’? with the same three answer boxes. Assume you take yourself to be a borderline case of baldness. The way the supervaluationist will diagnose the problem is quite like the ambiguous case; you are unsure how to judge whether you are bald, as the extension of ‘bald’ is unsettled between several candidates, and whether you are bald or not varies between these interpretations. Just as you could not judge if you had been to bank, you cannot judge if you are ‘bald’, as you cannot determine its full precise extension. As you tick ‘can’t tell’ for the first question, the supervaluationist should endorse ticking ‘can’t tell’ for the second question. It’s not right to tick ‘yes’ or ‘no’; you shouldn’t disbelieve you are bald, we simply can’t determine which property ‘bald’ is meant to pick out and thus determine whether you are or not.

The potential similarities between the semantic treatment of ambiguous and vague sentences extends past this. Suppose question three and four are the following:

3. Have you either been to the bank today or not been to the bank today?

4. Are you either bald or not bald?

As Fine points out, we can accept instances of LEM even for ambiguous sentences (1975:285-5). This is down to an analogous reasoning to the supervaluationist. You can

unproblematically endorse that you have either been to the bank or not, even in the questionnaire case in which you don't know how 'bank' is intended to be disambiguated, because LEM is true of both disambiguations. It's true that you went to the bank or not, whether 'bank' is meant to refer to the financial institution or the river side. Thus, even though you have been to the riverside and not the financial bank, and ticked 'can't tell' for question one, you can happily tick 'yes' for question three. This is of course very similar to how the supervaluationist sees question 4. In this case, even though you cannot determine which of a range of extensions 'bald' refers to, you can unproblematically accept that you are either bald or not. Even though if you tick 'can't tell' for question 2, you can happily tick 'yes' for question 4. This indicates that for both disambiguated and vague sentences, we can accept LEM, even though we don't assent to either of the disjuncts.

Something similar might also be said of sentences containing 'definitely'. Supervaluationists hold that an object definitely satisfies a predicate iff it is true it satisfies that predicate on every admissible precisification. You are only *definitely* bald if it's true you're bald on every admissible precisification. There isn't a straightforward counterpart of the 'definitely' operator when it comes to ambiguous sentences, so it isn't entirely clear, but still you can see a similarity; it seems it'd only be appropriate to apply 'definitely' to an ambiguous sentence if you know it's true on all disambiguations. Consider questions 5 and 6:

5. Can you tell me you've definitely been to the bank today?

6. Can you tell me you're definitely bald?

As we know the supervaluationist must endorse you ticking 'no' for question 6; being a borderline case, it's-- not the case that on every way of admissible precisifying 'bald' it comes out true that you're bald, it's straightforwardly false that you're definitely bald. It seems rational to do something similar with question 5. You cannot tell the questionnaire designer that you've *definitely* been to the bank; that would suggest to her that whatever disambiguation she has in mind, it's safe for her to believe that it's true you've been there today. Accordingly, as there is a disambiguation on which it's not true you've been to the 'bank', you must tick the 'no' box for 5.

There are clear similarities between ambiguity and the supervaluationist semantic treatment of vagueness, particularly when it is accepted that vagueness is semantic indecision. It is compelling that we should respond the same way to questions 3 and 4, and to questions 5 and

6. We thus might be compelled to respond the same way to questions 1 and 2. If we suspend on ‘I have been to the bank today’, when we don’t know how to disambiguate ‘bank’, there is an argument that the supervaluationist should suspend on ‘I am bald’, where we are a borderline case of baldness.

3. Conclusion

I’ve argued that supervaluationism does rule out some cognitive models but remains consistent with more than one option, and thus I conclude that there is a medium level of commitment between logico-semantic theories and theories of mind, as suggested in the previous chapter. I’ve also presented philosophical reasons for preferring a suspension model from the perspective of supervaluationism, it fits most cleanly with the logic and semantics, and the philosophical story commonly associated with supervaluationism.

In the following chapter I will discuss degree theoretic semantics and two candidate theories of mind that might be associated with it. In chapter 7, I will return to consider what the upshot of the conclusions in this chapter and the next would be for supervaluationist and degree theoretic interpretations of empirical data.

Chapter 6 Degree Semantics and Partial Belief

Degree theoretic semantics gives models on which truth comes in degrees. These models offer popular ways to handle vagueness in natural language and to solve the sorites paradox. Here I want to discuss what sort of constraints a degree theory places on an associated theory of mind. This involves assessing the cognitive significance of using degrees of truth, which I think is a slightly more complicated task than the supervaluationist case discussed last chapter. I will argue below that a logico-semantic model making use of degrees of truth requires a cognitive model which allows for degrees of belief. However, there is more than one way of giving such a cognitive model, and here I'll consider two different possible accounts, one from N. J. J. Smith, and one from John MacFarlane.

The upshot of these theories is that the rational attitude towards a borderline case (which will be true to a degree between 0 and 1) is a partial belief, somewhere between disbelief and full belief. A complication that arises for such a model is how these partial beliefs interact with partial credences; partial beliefs induced by uncertainty of the facts. I'll examine how both the models deal with this issue, and problems they individually face. According to Smith's theory, belief is a precise value, representing expected truth value. According to MacFarlane's theory, partial beliefs are best represented by a *probability distribution*. There are two ways we can interpret these as representative of attitudes; either as a kind of complex belief or a kind of indeterminate belief. I'll argue that the former doesn't give a genuine distinct option from Smith, but that the latter can give a viable alternative for the degree theorist.

Both these accounts are perfectly consistent with a degree theoretic approach to vagueness, but I will ultimately argue that we should prefer MacFarlane's theory. Smith's has the consequence that agents have a precise degree of belief in their attitudes towards borderline cases, and may also entail that different agents will have the same degree of belief as each other. This is less attractive than MacFarlane's picture, where agents have an indeterminate kind of belief, which can vary between agents, and between occasions for the same agent.

1. Degrees of Truth and Partial Belief

Degree theoretic semantics offers what some consider a promising way to deal with the problems of vagueness, by allowing that truth comes in degrees. Part of the appeal of this is

the way the theory can handle borderline cases. Degrees of truth allow us to model the apparent not-definitely-red-not-definitely-not-red patches neatly, by proposing such cases take a degree of truth between 1 and 0 (1 being full truth, 0 being full falsity). Thus, it might be true to degree 0.5 that a red/orange borderline colour patch, is red. This is meant to fit with our intuition that borderline cases of a predicate don't clearly fall into the positive or negative extensions of that predicate. It also fits with our intuition that we can distinguish between some borderline cases and how close they are to either the positive or negative extension of a predicate (which isn't as straightforward on a paracomplete model for example), as whilst two borderline cases can fall in a range between truth and falsity, the model can represent that one is closer to clear truth than the other.

Standard degree theories link back to Zadeh's (1965) work on fuzzy logic, but details between theories can vary. An important distinction is between truth-functional degree theories, and non-truth-functional ones. On a truth functional account, e.g. of the sort found in Machina (1976), the truth value of complex sentences is a function of the truth values of the atomic sentences that form it. For example, the truth value of a conjunction is taken to be equal to the truth value of which ever conjunct has the lower truth value. If P was true to degree 0.5 and Q was true to degree 0.4, (P & Q) would be true to degree 0.4. Degree models can be constructed that aren't truth-functional. For example, Edgington (1996) argues for a non-truth-functional degree account of truth, which she argues has exactly the same logical structure as probabilistic degrees of belief. Edgington also claims that her account is not a *degree supervaluationist* account (although it's not clear it is importantly distinct), which is another non-truth-functional account that can mirror the logical structure of probabilistic degrees (e.g. Williams 2011). The difference between truth-functional and non-truth-functional accounts are important, as we'll see, because of their connection to probabilistic belief. As I'll discuss below, we are going to need to know how our cognitive theory accommodates probabilistic partial belief. Here I'm going to be focussing on the truth-functional sort of degree theories, but it is worth keeping in mind how non-truth-functional theories might differ in their treatment of the issues at hand.

The question I want to address here is what sort of cognitive model best goes with degree semantics? And the related question, what sort of attitude is rational to take towards a partially true sentence? Recall that a complete theory of mind is a set of claims about rational belief and decision making, as examined in chapter 4. In this chapter, I am primarily

focussing on the consequences of introducing degrees of truth for rational belief. How are we going to model attitudes towards truth, falsity and partial truth?

If whilst browsing in a florist I stumble upon a borderline yellow French marigold, what attitude would a degree theorist think it is rational to take towards 'this French marigold is yellow'? As discussed in the previous two chapters, I think that accepting a particular logic and semantics places some restrictions on our theory of rational belief. In the case of degree theoretic semantics, I think the cognitive model should include a notion of *partial belief*. Exactly how we should understand partial belief is part of what is at stake in this paper, but it is commonly thought of as degrees of belief. I suggest that degrees of truth in a semantic theory mean that there should also be *something* like corresponding degrees of belief in the cognitive theory. The suggestion is that if it is partially true that the marigold is yellow, I may partially believe that the marigold is yellow, take an attitude between full endorsement and full disbelief. Just as truth comes in degrees, so can belief, being given a value between 1 (full belief) and 0 (disbelief); I may believe the 'French marigold is yellow' to degree 0.5 or 0.6 or 0.526. I believe that this is the route that degree theorists should take.

Recall from chapter 4 that what a model labels as 'truth' and 'logic' has significance for the cognitive implications of the model. Regardless of which semantics or logic we are working with, belief always *aims* at truth, the best that a belief can do is match up with truth. As Mark Sainsbury writes: 'where vagueness is at issue, we must aim at a degree of belief that matches the degree of truth, just as, where there is no vagueness, we must aim to believe just what is true' (1995: 44). The idea is that in a bivalent system or situation, the ideal belief states are full credence for true proposition, and zero credence for false propositions. When we extend the bivalent picture to a case where there aren't just two truth values, but infinitely many degrees of truth, we can keep the principle that belief aims at truth but extend the cognitive picture to one where there are more ideal possibilities than full credence and zero credence; the ideal degree of belief in a proposition is one that corresponds to its degree of truth.

As argued in chapter 4, I think that when a logic or semantics denies the connections between truth and logic and rational belief, the role of those notions becomes obscure. To quickly recap why, consider a degree theorist that favours a cognitive model on which the only rational attitudes are still belief and disbelief. We still ideally believe true propositions, disbelieve false propositions, but also disbelieve any proposition with a degree of truth less than one. On this picture we would disbelieve that the borderline yellow French marigold is

yellow, whatever degree of truth the proposition has, as it is true that it is yellow to a degree less than 1. This cognitive theory could be the same as one held by other different theories of vagueness, a three valued logic for example. On a three valued system, all borderline cases can be considered indeterminately true, and accordingly it might be suggested that the ideal attitude towards them is disbelief. Now we have a case where the degree theorist and the three valued logician have the exact same rational attitudes towards the same propositions, the only difference is that the degree theorist can distinguish between the degrees of truth of the borderline cases. But it's no longer clear what the importance of this is. The degree theorist here can *say* there are degrees of truth, but on what grounds do they deserve the label 'truth'? As also noted in chapter 4, it can't just be that these degrees are metaphysically significant properties. The three valued logician could also acknowledge the same degrees, but say call them instead 'degrees of closeness to truth'. It seems in this case the three-valued logician and degree theorist are disagreeing only verbally, about how to label these properties. For these degrees to deserve the label of 'truth' it should be that they indicate something significant for the cognitive role; that the rational attitude towards them is a partial belief related in some way to that degree of truth.

This connection between degrees of truth and degrees of belief is seen in the literature also, N. J. J. Smith takes it for granted that proponents of degree theoretic accounts of vagueness should already believe that vagueness also induces degrees of belief (Smith 2010, in Dietz and Moruzzi 2010: 491), and similar can be seen in Machina (1976) and MacFarlane (2010). The models I'll be considering here take the rational attitude towards borderline cases to be a kind of partial belief.

In inducing partial beliefs, a question quickly arises for these models; what is the connection between these beliefs and *probabilistic credences?* Probabilistic partial beliefs are those beliefs which we do not full endorse due to being *uncertain*. We are unsure about a proposition's truth value due to not being certain of the facts. It is common to model our credence in propositions of which we are uncertain with degrees also. For example, I believe that my mother would like some nice flowers for Mother's Day, as her well-tended garden suggests she likes flowers, but I don't have enough evidence to be certain that she would. I have some degree of confidence that she would, which falls short of absolute certainty. Similarly, I don't believe that she would like a Van Halen CD, as she doesn't often listen to 80's rock bands, but again I am not absolutely certain. Here I have some degree of confidence that she wouldn't like a Van Halen CD, which falls short of absolute certainty. We might

represent these levels of confidence as degrees of belief, taking a value between 1 (representing absolute certainty) and 0 (representing absolute certainty of the negation). My degree of belief that my mother would like flowers might be quite high and close to certainty, perhaps 0.8. Correspondingly, my degree of belief that she would not like flowers would be 0.2. My degree of belief that my mother would like a Van Halen CD might be very low, say 0.1, and my degree of belief that she would not like the CD correspondingly high at 0.9. This is a standard approach to thinking about credence in vagueness free situations (pioneered by Ramsey, 1926), proposing belief comes in degrees in a *probabilistic* framework; our credence in a sentence reflects how likely we take the truth of that sentence to be.

The probabilistic degrees of belief are distinct from partial belief induced by vagueness. A flower may have petals of a colour which is somewhere in the region between pink and purple, and thus be a borderline case of a pink flower. If asked if such a flower is pink, I may lack confidence in an answer which affirms it is or that it is not. I may have some partial confidence that it is a pink flower, which falls short of my confidence that clear cases of pink flowers are pink. Similarly, I may have some degree of partial confidence that it is not pink, which also falls short of my confidence that clearly not pink flowers are not pink. Unlike the prior case, this is not uncertainty due to lack knowledge about the facts. I can know everything about the colour of the flower, as well as the use and meaning of the word 'pink', and still not fully endorse that the flower is either pink or not pink. This too can be modelled with degrees of belief, but this leaves us asking how to two sorts of partial belief interact.

The degree theorist is going to need to be able to accommodate both sorts of partial belief considered here, and thus must be able to say what happens in cases where we are affected by both uncertainty and vagueness. Say I decided to buy my mother flowers, thinking it's a safer bet than a Van Halen CD. I get to the shop and have to decide what colour flowers to buy. Last week my mother showed me her favourite flower, which was a pinky-purpley colour; it was borderline pink. In order to help with my decision of what flowers to buy, the florist wants to find out what colour my mother's favourite is, and asks 'is your mother's favourite flower pink?' How do I answer? Specifically, what attitude do I take to the sentence 'my mother's favourite flower is pink'? This sentence is affected by the vagueness of pink; as the flower she showed me was borderline pink, something doesn't seem quite right about calling it pink, but something also doesn't seem quite right about saying it isn't pink. As outlined above, in this case the vagueness of 'pink' might provoke us to say that it is true to a degree between 1 and 0 that the flower is pink. The belief about whether the colour of my mother's

favourite flower is pink is also affected by general uncertainty. I may now not be entirely certain that the flower my mother showed me last week is still her favourite; perhaps she has since chosen a new favourite flower, which is not pink. I know for example, that around this time of year, when spring has just begun, my mother takes a strong liking towards Daffodils, which are definitely not pink. This lack of certainty about the facts might also provoke us to say that I have a degree of belief that my mother's favourite flower is pink which is less than 1 and greater than 0. How do these two different sorts of partial belief interact? I seek a theory here that can provide some answer to that question.

So, I think a successful cognitive theory for the degree theorist should include a notion of partial belief, and it should tell us how this sort of partial belief interacts with probabilistic credence. I know turn to consider two theories which can give some of the cognitive theory for degree semantics. An important aim for both was to be able to combine partial beliefs induced by vagueness and uncertainty. They differ more greatly on their conception of partial belief, with Smith assigning precise degrees of belief to subjects, and MacFarlane allowing for a slightly blurrier picture by use of probability distributions.

2. Smith: Belief as Expected Truth Value

N. J. J. Smith (2010, 2014) gives a theory of rational belief for degree theoretic semantics, according to which belief in a proposition is a weighted average of its possible degrees of truth. At least three important features of the model can be picked out:

1) Rational belief is a *precise* degree of belief

2) Degree of belief is expected truth value

3) Degrees of belief can be understood as strength of tendency to act.

The first important point to note is that on Smith's model, belief comes in degrees, and rational belief in a borderline true proposition is a degree of belief corresponding to its degree of truth. The cognitive significance of 'the flower is yellow' being true to degree 0.5 is that the rational degree of *belief* in the flower is 0.5. So, an important part of this model is that rational belief isn't an all or nothing attitude, where there is vagueness at stake, and the truth of a proposition can vary in degrees between 1 and 0, belief in that proposition can also vary. This may be a plausible way to deal with vagueness; it might be argued that as we pass through a sorites series from say clear yellow cases to borderline cases to clearly not yellow

cases, our confidence in the applicability of 'yellow' gradually drops as the objects in the series gradually become less yellow.

So on this model, vagueness can cause rational belief to fall to a degree between 0 and 1, but as discussed above, *uncertainty* is another reason we might model belief as a value between 0 and 1. If you put a flower behind your back, and I don't know whether it's in your left or right hand, rational belief in the proposition 'the flower is in your left hand' will also be a degree between 0 and 1, in this case 0.5 (as there is half a chance it's in your left hand and half a chance it's in your right). In cases where vagueness is not at issue, but there is uncertainty, this is the result given by Smith's model.

When it comes to a proposition which is affected by both uncertainty and vagueness, Smith suggests we use a formula to generate a single degree of belief. For example, say you take a borderline yellow flower and put it behind your back, and again I don't know whether you are holding it in your right or left hand. What is my degree of belief in 'there is a yellow flower in your left hand'? Here I don't fully endorse that the flower is yellow, and I don't fully endorse that it is even in your left hand. Smith suggests we can combine these partial beliefs by taking a weighted average of the degree of truth of the proposition across a set of possible worlds. In the current example there are two possibilities, the flower is in your left hand (world 1), or the flower is in your right hand (world 2). Say that the borderline yellow flower is yellow to degree 0.5, we would get the following result:

truth value of proposition at world 1 x likelihood of world 1 + truth value of proposition at world x likelihood of world 2 =

0.5 (as flower is 0.5 yellow) x 0.5 (as there is a 50% chance of it being in your left hand) + 0 (as proposition is false when flower is in the right hand) x 0.5 = 0.25 + 0 = 0.25

So I believe the proposition 'there is a yellow flower in your left hand' to degree 0.25, according to Smith's model.

Let's take a more formal look at the model. For Smith, our degree of belief in a proposition is the average truth value of that proposition at a range of possible worlds (the ones the agent has not ruled out) weighted by the likelihood we take in each world being the *actual* world (2010, in Dietz and Moruzzi 2010: 502). Formally, Smith takes an agent's epistemic state to be a function *P*, which maps the real numbers from 1 to 0 to subsets of possible worlds. The number assigned to a subset of possible worlds represents how likely the agent thinks it is

that the actual world is amongst those in that subset. At each possible world, a proposition *S* determines a function S', which assigns the proposition a real number between 1 and 0. This represents the degree of truth that proposition has at that world. Smith proposes that an agent's degree of belief in proposition *S* is determined by these values, to generate the expected truth value of S:

$E(S) = P(\{w1\}) \times S(w1) + \ldots + P(\{wn\}) \times S(wn)$

(Smith 2010, in Dietz and Moruzzi 2010: 496-7)

We can see how this gives the results described above. Where uncertainty isn't at issue, but vagueness is, belief matches degree of truth, as $P(\{w1\})$ will equal 1, and there is only one world at issue. For instance, if I see a borderline yellow flower, which is yellow to degree 0.5, my degree of belief in proposition *S*, 'this flower is yellow', is as follows:

$E(S) = P(\{w1\}) \times S(w1) = 1 \times 0.5 = 0.5$

Where vagueness isn't at issue, but uncertainty is, degrees of belief will act the same as probabilistic credences, as *S(wn)* will only be 1 or 0. For instance, when you hide the flower behind your back, and I cannot tell if tell if it is in your left or right hand, my degree of belief in proposition *S*, 'the flower is in your left hand', will be as follows (given there is a 50/50 chance of the flower being in you left hand (*w1)* and your right hand (*w2)*):

$E(S) = P(\{w1\}) \times S(w1) + P(\{w2\}) \times S(w2) = 0.5 \times 1 + 0.5 \times 0 = 0.5$

So, we can see the first two features of the model I outlined at the beginning of this section. First, rational belief is a precise degree of belief: whatever the levels of uncertainty and degrees of truth of the relevant proposition, belief in that proposition always comes out as a particular degree of truth. Second, degree of belief is expected truth value: an agents belief in a proposition is its average truth value across possible worlds that the agent hasn't ruled out (that may be the *actual* world) weighted by how likely each of those worlds is taken to be.

The third important feature is that Smith understands these degrees of belief as *strength of tendency to act*. There is an inherent issue when it comes to combining degrees of belief from uncertainty and from vagueness, given that the prior is probabilistic, but the latter is not. We have to ask Smith how we are meant to understand these degrees of belief, what it means to believe a proposition to degree 0.6 on his model. It is easy to understand a probabilistic credence, as it is usually well explained through the idea of betting behaviour: our degree of

belief of 0.8 in 'P' can be well explained by how often we would bet on 'P' being true. This sort of explanation of a degree of truth cannot be straightforwardly applied to a degree of belief in a proposition when it is affected by vagueness. Furthermore, as degrees of belief invoked by vagueness do not conform to probability calculus (on Smith's model), it can be argued that an agent would be subject to a *Dutch Book*: that is, to rationally placing a bet where they are guaranteed to lose. Smith suggests that it would always be irrational to place a bet at all where vagueness is concerned, as the outcome would be undetermined (2010, in Dietz and Moruzzi 2010: 503). For Smith, in the above scenario I couldn't rationally place any bet on there being a yellow flower in your left hand, given that I believe it was borderline yellow, as it is *indeterminate* whether it is yellow, and thus indeterminate whether I win or lose the bet. That all betting behaviour is irrational where vagueness is concerned allows for Smith to avoid the threat of a Dutch book[8]. Without betting behaviour to give a deeper explanation of the degree of belief, Smith must offer an alternative.

Smith suggests that a degree of belief in proposition *S* should be interpreted as the strength of tendency to act as if *S*. He claims that this has an advantage over explanations in terms of betting behaviour, as it can be applied to situations affected by both vagueness and uncertainty, and situations where there is no vagueness at all (2010, in Dietz and Moruzzi 2010: 505). Smith takes this understanding of degrees of belief as a general motivation for his framework. He suggests we can only one strength of tendency to act as if *S* in any given situation. One cannot simultaneously have a strong tendency to act is if my mother's favourite flowers are pink (thus buying her lots of pink flowers for Mother's Day) *and* a weak tendency to act as if my mother's favourite flowers are pink (perhaps buying some pink, but also some yellow, or no pink flowers at all). I can only act in one way, according to one strength of tendency to act as if *S*. Smith argues this is motivation for having a *unified* notion of degree of belief. We cannot allow that our degrees of belief invoked by uncertainty and by vagueness remain separate, as then we might have two different degrees of belief in some proposition, and thus two different strengths of tendency to act. (2010, in Dietz and Moruzzi 2010: 494). This fills in part of the story, but I don't think it's entirely clear what strength of tendency to act really means, which I'll discuss in the following section.

[8] As Smith is aware (e.g. 2014: 1038), one option is to extend the notion of betting to include part degrees of truth and partial pay outs, but it isn't necessary to discuss this here. See Paris (2001) and Mundici (2006) for further discussion

We now have one model of rational belief for degree semantics. Partial beliefs are given as precise degrees of belief, which is our expected truth value of the relevant proposition. These degrees of belief correspond to strengths of tendency to act. The way Smith handles our degree of belief in a proposition seems quite peculiar, as in his model our degree of belief in proposition is a function of the truth value of a proposition at a world, not of our degree of belief in that proposition at a world. It appears that in doing this, Smith is suggesting that our degree of belief in a vague proposition will match its precise degree of truth. That when I see the 0.5 yellow flower, my degree of belief that the flower of yellow will automatically match its degree of truth. If this is Smith's intention, I think it is an implausible feature of a cognitive theory. I will discuss whether it is his intention below. We might also worry whether degrees of belief can be very straightforwardly understood as strengths of tendency to act. In the following section I'll consider these issues, and Rosanna Keefe's discussion of them, concluding that Smith's model may at least need revision on these fronts. I also worry about the precise degrees of belief assigned to participants, I am not convinced that subjects attitudes towards borderline cases can be well represented by a single particular degree of truth.

2i. Evaluating the model

Keefe (2016) has two important objections to Smith's proposal. One is that the model takes degree of belief to be a function of both subjective judgements based on evidence and objective truth values in the world, which Keefe argues is an inappropriate role for objective truth to play in a psychological model. A second is that Keefe doubts that there is a clear link between action and degree of belief in cases where vagueness is concerned, relating to concerns from Dorothy Edgington (1996). I don't find doubts about the connection between degree of belief and action convincing, and will argue that there is no obstacle to Smith's theory (or another theory which aims to connect the two, such as MacFarlane's (2010), to be considered in the following section) to be found there. However, I do agree that the role of truth values in the model seems inappropriate. Whilst I don't fully endorse Keefe's worries about their role in a psychological model, I do agree in this case they lead to the undesirable consequence that different agents will always have the same precise degree of belief in vague propositions.

I'll consider the connection between degree of belief and action first. Keefe is critical of Smith's explanation of a unified degree of belief as strength of tendency to act. Keefe thinks

that the connection between belief and action is not sufficiently clear for this explanation to work, at least where vagueness is concerned, and Edgington has similar worries about this link.

Smith's proposal is that a degree of belief in a proposition *S* can be explained as the strength of tendency to act as if S. The greater our degree of belief in *S*, the stronger our tendency to act as if *S*. This does seem to work out in some cases, as we can see from an example of Keefe's (2016: 3796). When picking players for a basketball team, if you want someone tall and are faced by two borderline tall men, x and y, you are more likely to choose x than y, if x is taller (all other things being equal). Smith might claim that this is because your belief that 'x is tall' is greater than your belief that 'y is tall'. On this model, I have a strong tendency to choose the players that I fully believe are tall. As we move through a sorites series of people from the definitely tall through the borderline cases, my belief that each is tall slowly decreases, and with it does my tendency to choose each. Eventually we reach those people which I fully believe are not tall, and accordingly I have no tendency to choose them.

Yet Keefe and Edgington doubt that this sort of link is always so straightforward. In some cases, our degrees of belief might not always line up with actions. Imagine that I am in a florist, to buy my mother some flowers for Mother's Day. Suppose that I check in advance, and my mother tells me her favourite flower is currently the daffodil, a flower that is definitely yellow, so I decide that I want to buy some yellow flowers for her. Say I get to the florist, but unfortunately find that they have no daffodils. The two best candidates for a yellow flower are a bunch of French marigolds which are borderline yellow, with an orangey tinge, and some amber flush roses, which are also borderline yellow, but are slightly closer to orange than the marigolds. As follows from Smith's model, I may have a degree of belief that the marigolds are yellow that is greater than my degree of belief that the roses are yellow. But in this case it may not translate into action. I may not be more likely to buy the marigolds than the roses; I may be equally unlikely to buy either. After all I wanted *yellow* flowers, so my partial belief that the marigold or roses are yellow may not be enough for me to choose either. In this case my tendency to buy a flower may not steadily decrease over a sorites series, starting with a definite yellow daffodils and butter cups, through the French marigolds and amber flush roses, to a clearly orange gerbera daisy. I may have tendency to buy a definitely yellow daffodil or a butter cup, but no tendency at all to buy a borderline yellow marigold.

In a similar vein, Edgington points out that I may have a preference for coffee over tea, but my tendency to reach for a cup my not decrease steadily over a sorites series of mixtures, from definite coffee, though indeterminate cases of coffee/tea, to definite tea. I may be more likely to reach for tea, than for the borderline mixtures of coffee and tea (1997, in Keefe and Smith, 1997: 313), even though I have a preference for coffee, and believe tea is coffee to degree zero.

As Smith suggests, it appears one way to try to avoid these counter examples is by being more careful about the description of our desires in each case (2010, in Dietz and Moruzzi 2010: 498). It could be objected that we in fact *wouldn't* choose a borderline tall person for our basketball team, if we wanted *tall* people. In this case, Smith argues that this is because we want 'very tall' people, and so borderline tall people may not fit the bill; we would believe that borderline tall people are 'very tall' to degree zero. I'd prefer to put the point using the definitely operator, than using a modifier like 'very', as I think it better captures what is likely to be going on in the above counter examples. In the basketball case, we may actually want *definitely* tall people. If we did, we wouldn't be likely to sign up a borderline tall person to our basketball team, as we would believe they are definitely tall to degree zero.

It might be argued we can avoid worries in the flower and coffee case by re-describing desires also. It's true that if I wanted to buy yellow flowers for my mother, the strength of my tendency to buy a flower may not smoothly decrease over the sorites series; as in the above example, it may go from a strong tendency to buy a daffodil, to no tendency to buy a borderline yellow French marigold. However, in such a case, the issue could be resolved with a different description of the desire. On the sorites series, if my tendency to buy a flower stops at the definitely yellow cases, and I have no tendency to buy a flower of which is not definitely yellow, this is because my desire is not for a yellow flower, but for a *definitely* yellow one. Similarly, we can dispel the coffee counter example by correct description of desires. I may not have a tendency to reach for a cup which smoothly decreases over a sorites series from coffee to tea, as my belief that each cup is coffee does. I may have a strong tendency to reach for a cup of something which is definitely coffee, but not tendency to reach for a borderline coffee/tea mixture. This is because I desire something which is definitely coffee, and have no desire for something which is a mixture of tea and coffee. We can also explain why I could be more likely to have a cup of tea over a coffee/tea mixture. I may have a weak desire for a cup of tea, but no desire for something which is borderline coffee/tea. This gives something of a plausible way to deal with the coffee case.

Keefe is not convinced that such a defence can always work. Whilst she accepts that in the coffee/tea case, we *can* avoid worries by explaining we have a desire for a mixture which is definitely coffee, she thinks others are imaginable which would not be so easily evaded:

> For example, consider cases where there is a less gradual drop-off in the extent to which my desire is satisfied by borderline Fs, corresponding neither to the degree of "x is definitely F" nor of "x is F". Such a pattern can be perfectly natural and rational. (Keefe 2016: 3796)

Here the suggestion is that while in some cases our actions can be explained by our desire for something definitely F, and our degree of belief that certain objects satisfy definitely F, or by our desire for something F, and our degree of belief that certain objects satisfy F, this does not explain all rational action. In some cases our tendency to act as if Fx may extend past the cases of definitely F's, but not decrease smoothly through the borderline cases, as does my degree of belief that each x is F. Rather, it may extend past the definite cases, and drop off suddenly part way through the borderline cases.

Back in the florists again, where I'm still searching for yellow flowers, Keefe is suggesting that I may have a strong tendency to reach for a clearly yellow buttercup flower, but that my tendency to reach for a flower may not smoothly decrease over the other borderline yellow flowers in the shop. It may drop off suddenly; say between the French marigolds and the amber flush roses. Keefe suggests that such behaviour would be completely natural, yet it cannot be so easily avoided by distinguishing between a desire for a yellow flower, and for a definitely yellow flower, given the marigold and roses are both borderline cases.

Can Smith reply to this? I think we can put pressure on the suggestion that such behaviour pattern really is natural and rational, where the only relevant considerations are belief and desire. If when choosing flowers, the only thing in play is my desire for yellow flowers, and the degree of belief that each is yellow, I'm not entirely convinced the sort of pattern Keefe is considering would be that natural.

Let's say I decide to buy every yellow flower in the florist. The florist has conveniently laid out a series of ten flowers, the first of which is yellow to degree 1, the second to degree 0.9, and so forth, to the tenth which is yellow to degree 0 (and these are the ten closest flowers to yellow in the shop), and in this instance my degree of belief that each is yellow matches the degree of truth to which each is yellow. On Smith's model, the strength of my tendency to buy each flower slowly decreases over the series. I will certainly buy the first flower, which is definitely yellow, but become gradually less likely to buy each flower as they become

slightly less yellow. However, I still have a slight tendency to act as if the ninth flower is yellow (believing it is yellow to degree 0.1). Keefe is suggesting that it would entirely natural in cases like this to have a tendency to buy the flowers that are yellow to a high degree, but to have no tendency to buy the flowers which are yellow to a low degree at all. I may be very likely to buy the first three, somewhat likely to buy the next three, but have no tendency to buy the final four at all.

I am not convinced that this kind of pattern is so natural, at least where all things are equal except for the simple desire for yellow flowers. If this my only relevant desire, it is plausible to take it that this desire is well satisfied by definitely yellow flowers, and gradually less satisfied by each as they become gradually less yellow. None the less, the seventh flower, being yellow to degree 0.3, still offers *slight* satisfaction of this desire, as do the eighth and ninth. I still have a weak belief that these flowers are yellow, and thus I'm not sure it's entirely natural that I'd have no tendency at all to act as if they were yellow (i.e. buy them in this case).

I'm not suggesting that this means that in every instance I will buy all flowers that are yellow to any degree at all, I might entirely rationally not buy *all* nine flowers. But Keefe's objection is not that one could rationally refuse to buy the flower that is yellow to degree 0.1, rather it is that I may have no strength of tendency to act as if it is yellow at all. This is the claim that I dispute; I think I will have some tendency to act as if it were yellow, given that I do have some weak belief that it is yellow. I think that any case where my strength of tendency to act as if each flower is yellow *does* drop off say between the sixth and seventh, and I have no tendency to act as if the seventh flower is yellow at all, indicates that not all other things are equal. If I have *no* likelihood at all of buying a flower I believe to be yellow to some degree, then I think it's likely that I don't have a simple desire for yellow flowers anymore. Perhaps I have a desire for *quite* yellow flowers, or *moderately* yellow flowers, which the seventh flower onwards cannot satisfy. After all, the sixth and seventh flower, being yellow to degree 0.4 and 0.3 respectively, are both borderline cases of 'yellow'. If I had no strength of tendency to treat them as yellow at all, then they'd be no different to definitely orange flowers, and this doesn't seem to acknowledge the significance of them having a degree of truth above 0. It may well be the case that I decide not to buy the seventh, but I don't think it would naturally be the case that I have no tendency to buy it at all, where my only relevant desire is for yellow flowers.

Of course, how exactly I behave depends on how different *strengths of tendency* to act play out, which is something that Smith hasn't fully made clear to us. The notion of a strength of tendency to act is also invoked to help understand probabilistic credences, as seen in Ramsey:

> the degree of a belief is a causal property of it, which we can express vaguely as the extent to which we are prepared to act on it. . . . it is not asserted that a belief is an idea which does actually lead to action, but one which would lead to action in suitable circumstances . . . The difference [between believing more firmly and believing less firmly] seems to me to lie in how far we should act on these beliefs. (1990: 65-6)

But in this case we can expand this notion slightly with the idea of betting behaviour: a 0.5 degree of belief in *S* represents a level of confidence such that the agent would be willing to bet a stake of 50% of the possible return (e.g. bet 50p for a return of £1). However, Smith thinks that this cannot be used in the current case, as it wouldn't be reasonable to make a bet where there can be an indeterminate outcome. In the case where I am unsure which borderline yellow flower you are holding behind the curtain, my 0.5 degree of belief that you are holding a yellow flower, cannot be associated with a degree of confidence such that I would bet 50p that it will be yellow for a £1 return, in Smith's opinion. As the flowers are borderline cases, it wouldn't be possible to determine if the flowers are yellow or not, and thus not possible to determine if I won the bet or not.

Smith doesn't offer a particularly explicit explanation of the notion, but has in mind that we will still behave with varying levels of confidence. For instance, he notes that depending on your degree of belief that Fido (the dog) is dangerous, you might either tremble or offer him some beef jerky when he enters the room. How would this pan out in the case of wanting to buy yellow flowers in the florist? For the flowers I believe to be yellow to a high degree (but below 1), I act *strongly* as if they are yellow, perhaps deciding to buy them very quickly. For the ones that I believe are yellow to a low degree, I act *weakly* as if they are yellow, perhaps considering them briefly as candidates, before deciding not to buy them. Smith also notes, as already seen with the case of choosing 'tall' basketball players, that a higher degree of belief that one person is tall that another will result in a stronger tendency to pick that person.

An alternative possible analysis of varying strengths of tendency to act, corresponding to degree of belief, is in terms of *frequency;* a degree of belief in proposition *S* could be thought to represent how likely you are to act as if *S* is true, so the percentage of times you would as if *S* is true, rather than false. In the case where I want to buy every yellow flower in the florist, the degree to which I belief each flower is yellow may correspond to the frequency at which I buy each flower (or equivalently, the chance of my buying them). Believing that the

daffodils are yellow to degree 1, I will certainly buy them, believing that the red roses are yellow to degree 0, there is no chance of me buying them. Across the ten flowers laid out, which decrease from degree 1 to degree 0 at equal intervals, I would by the first every time, the second nine times out of ten, the third eight times out of ten, and so forth.

This suggestion isn't entirely *ad hoc*. For one thing, Smith didn't give us a well-developed notion of tendency to act, and frequency looks like a reasonable possible analysis: as the strength of tendency to act as if *S* increases, the *likelihood* of acting as if *S* increases. Furthermore, an analysis of degrees of truth in terms of percentages of action has been suggested before. For instance, Max Black suggested this kind of analysis (1937). Black defines the notion of a *consistency profile* for the use of category, say *F*, which is the ratio of speakers that would apply *F* to some object, *a*, to speakers that would apply ¬*F*. He suggests that when it comes to vague language, this ratio will tend to a value far above one for clearly true cases, and far below it for clearly false cases, but will tend toward 1 for some cases between those (1937: 442-3). The cases for which the ratio would tend towards 1 are of course the borderline cases, and the ratio of 1 shows that half the speakers would apply *F*, and half would apply ¬*F*. He goes on to suggest that the consistency profile could represent the degree to which a category applies to an object:

> If the analysis of *L* [a vague category], (i.e. the specific consistency profile) be denoted as *L'*, an alternative mode of formulation would be to regard the consistency distribution as indication of the degree to which *L'*, the more explicit symbol is applicable to the corresponding terms of the series *S* [a sorites series]. (pp. 446)

The idea then is that the degree to which a vague word applies to an object is represented by its consistency profile. The degree to which *a* is *F* represents how likely speakers are to apply *F* to *a*. This taps easily into the suggestion that we can analyse *tendency to act* as a *frequency* of action. As the degree truth of a proposition rises, so does an agent's degree of belief that it is true, and their strength of tendency to act as if it is true. As suggested by Black, an interesting way to interpret this is as a greater likelihood to behave as if the relevant proposition is true. At the individual level, this means degree of belief in P means likelihood of acting as if P as a percentage (for a 0.5 belief, is 50% likely to act as if P). At the group level, this correspond to the percentage of people who act as if P[9].

[9] A similar idea is found in Egré (2017), which is based on the work of Borel (1907)

The former of these two ways that strength of tendency to act might play out is a little unclear, what are the effects of a 0.6346 belief that *S* on behaviour, compared to 0.3647? But the latter gives something straightforward to work with. For example, in the reply to Keefe's concerns about belief and action, we can say that the greater my belief that a flower is yellow, the more likely I am to buy it. If Keefe is concerned that someone might very reasonably choose not to buy the sixth through tenth flowers in our ten flower series, then that can be accommodated, as those flowers would be picked a low percentage of the time. I can rationally not buy the flower I believe to be yellow to degree 0.1, that belief means I would buy it only 10% of the time.

This sort of connection also gives the degree theorist a good avenue to connect with experimental literature and the experimental approach to vagueness. We can expect that the higher the degree of truth that a vague predicate applies to an object, the greater the percentage of subjects that do describe it so. In psychological and experimental philosophical literature, we can see patterns whereby the applicability of a vague predicate increases gradually over a sorites-type series, as gradually more subjects are likely to apply the predicate. Similarly, we see that subjects change their mind about whether a vague predicate is applicable or not when undertaking an experiment for the second time; with the current proposal, this is neatly explained by degree of belief representing the likelihood of applying the predicate, such that there are occasions where the subject would apply it, and occasions the subject wouldn't. I will go into more detail on how the cognitive models discussed here and in chapter 5 can connect with past and future studies in the following chapter.

2ii. Objective truth values in a subjective model

Keefe also objects to the use of truth values in Smith's model. For Smith, my belief that 'there is a yellow flower in your left hand' is determined by my *perceived* likelihood that the flower is in fact in your left hand, as well as the *objective* degree of truth that the flower was in fact yellow. Keefe is concerned that the model suggests that a psychological state is directly determined by objective truth values, rather than some sort of subjective perception of those truth values. Indeed, something does look a little strange about this. There are at least two good reasons Keefe has for objecting to it. One is that a *psychological* model shouldn't be take objective truth values into account, as agents might not be able to track them:

> This picture cannot succeed in illuminating our degrees of *belief* however. It allows, relative to each world, for a contribution from borderline cases reflecting the extent to which they are

> true—their actual truth-values—but thereby takes on an objective feature of the situation rather than the subjective state of the believer in question… But degrees of *belief* should correspond to our different *judgements* of, say, how close a borderline case of F-ness is to being F which might not be an accurate judgement of degree of closeness. (Keefe 2016: 3793)

A second is that a consequence of degree of belief being a function of truth values is that different agents will always have the same degree of belief in the same propositions (where uncertainty wasn't an issue), but this looks wrong:

> Whatever one's theory of vagueness, if we want a story about subjective degrees of belief in the light of vagueness, we need to allow for inaccurate subjective states and different states for different people. Bob may be more "generous" with his judgement of tallness than Cam. If forced to guess where the boundary was (or draw a suitable boundary) Bob would place it lower than Cam. And faced with Tek [a borderline case of tallness], Bob would be a little more inclined to call him tall than Cam would. It would be natural to say here that Bob's degree of belief in 'Tek is tall' is higher than Cam's. (2016: 3794)

So one worry about the role of truth values in the model is about whether they should even appear in a *psychological* model; agents may not reliably track precise degrees of truth, and perhaps a subjective notion, like *judgements* of a propositions degree of truth would be more appropriate. A second worry is that if truth values do play this role, then it will result in the agents always having the same degree of belief in vague propositions; upon looking at a borderline pink flower, surely everyone wouldn't have exactly the same degree of belief that that flower is pink. If this is meant to be a descriptive account of psychological states, shouldn't it allow for individual differences when it comes to attitudes towards borderline cases[10]?

Let's consider this in light of an example. So, according to the model, degree of belief is the *expected value*: average truth value of the relevant proposition across possible worlds weighted by how likely the relevant agent takes each world to be. Say you hide four borderline yellow flowers (which I've seen) behind a curtain, flower 1, flower 2, flower 3 and flower 4. Flower 1 is yellow to degree 0.8, flower 2 is yellow to degree 0.6, flower 3 is yellow to degree 0.4, and flower 4 is yellow to degree 0.2. You will select one at random and remove the three others, and hold it behind the curtain obscured from my vision. What is my degree of belief in proposition *S:* 'you are holding a yellow flower'? According to the model it is the following:

[10] Smith does make clear that this is the commitment of his view: 'In other situations, the agent will be free of uncertainty with respect to some propositions of interest: she is certain of exactly how true they are in the actual world' (2010, in Dietz and Moruzzi 2010: 502).

$E(S) = P(\{w1\}) \times S(w1) + P(\{w2\}) \times S(w2) + P(\{w3\}) \times S(w3) + P(\{w4\}) \times S(w4) =$

$0.25 \times 0.8 + 0.25 \times 0.6 + 0.25 \times 0.4 + 0.25 \times 0.2 = 0.5$

Degree of belief here is the expected value of *S*, which is 0.5.

It seems Keefe is concerned that the subject may not have reliable access to the degree of truth of *S* at these worlds, so it is strange that in the model they contribute to belief in the way they do. In this case, my degree of belief in *S* is a function of *S*'s truth value at worlds *w1-w4*, but Keefe doubts that I can access the degrees of truth in those four scenarios so accurately and precisely. Why think I'd take the degree of truth of flower 1 to be 0.8, rather than 0.79, or 0.81, or 0.8355? Keefe argues that the values given by the model cannot illuminate belief, unless agents can be thought to track these values. What's more, the model predicts *any* agent would get the same degree of belief in proposition *S* in the above scenario. But surely we can imagine speakers having some disagreement about just how close flower 1 or 2 is to yellow; particularly when it comes to borderline cases, speakers may well disagree about how well an object exemplifies a relevant vague category.

We don't need to be worried that the model takes belief to be a function on truth values. In general, this is unproblematic. For example, I have taken it for granted thus far that a competent speaker of English would be able to accurately apply 'tall' to determinately tall cases, someone seven-foot-tall for instance. If *S* is the proposition 'John is tall', where John is seven-foot-tall, we can expect such a speaker to fully believe that proposition. In this case, it wouldn't be problematic to take belief to be a function of the truth value of *S*: anyone who understands the meaning of 'tall' would fully believe proposition *S*. Consider the flower case above, but suppose that all of flowers 1-4 are clearly yellow daffodils. In this case, if you're going to pick one randomly, I would fully believe 'you are holding a yellow flower', because any competent speaker would fully believe that each of those flowers are yellow. If flowers 1-4 were red roses, then I would disbelieve 'you are holding a yellow flower'. Here there is no issue taking the truth values of these propositions as a reliable indicator of my beliefs; as a competent speaker, my beliefs track the truth in these cases. Talking about clear true and false cases gives us Smith's model, but restricted to truth values 1 and 0, and in this case the model is a good one (it is in fact just a classical probabilistic model of belief).

If we can accept that much, a step to Smith's full model, which generalises to degrees of truth, isn't problematic. We can expect competent speakers to track the truth in clear flower cases above, and Smith's model takes it we can expect competent speakers to track the truth

in borderline cases also. This is based on the idea I argued for in chapter 4 and the outset of this chapter, that a semantic model is committed to saying something particular about belief, and in the case of degree theoretic models, there is a commitment to saying that there are corresponding partial beliefs of some kind. Given this, it's right that an agent's attitudes to a proposition are a function of that proposition's truth value (in the way they are in the classical case); after all, on this picture, at least part of the significance of saying there are partial truth values is just that agents will have related partial beliefs. There is nothing wrong with taking beliefs of competent speakers to track truth values in the classical case, nor in the degree case. Accordingly, I reject Keefe's first reason for concern; there isn't anything wrong with objective truth values being used in a model of belief in general.

I do think the second concern, that Smith's model entails that everyone's degree of belief will be the same in the same propositions (under the same epistemic circumstances, i.e. levels of uncertainty), is more worrying. Whilst I think it's right that belief is a function of truth values, it is implausible that any agent will get the exact same degree of belief when faced with a borderline case. Degree of belief is taken here to be a *strength of tendency to act*, and whilst it certainly is plausible that speakers will act in similar ways relative to some borderline case, Smith's model gives the result that every agent will act in a way matching the *very precise* degree of belief they all share in the relevant proposition; as degrees of belief will precisely match truth values where there isn't any uncertainty in play. I do think this is an implausible result when it comes to borderline cases. Say there were a borderline yellow flower, *f*, which is yellow to degree 0.675. According to the model, every agent will believe *'f* is yellow' to degree 0.675, and have the same strength of tendency to act as if it is yellow, e.g. be equally likely to buy it from a florist when shopping for flowers. This seems wrong, it seems likely that some agents would be more likely to act as if it were yellow than others. Note the point isn't just that agents will act differently towards that flower, as Smith himself notes (e.g. 2014: 1029), agents with the same beliefs (and strength of tendency to act) may act in many sorts of ways depending on a variety of factors (their desires for instance). The point is that there should be room for *some* discrepancy between agents in their strength of tendency to act. If five people came in to the florist who wanted to buy every yellow flower they had, should we expect they'd all be equally likely to act as if flower *f* is yellow? It is plausible that give *f*'s borderline status, there would be *some* level of disagreement about just how close to yellow *f* is, with some people perhaps deliberating for a while about whether to buy it, perhaps some choosing to buy it, and perhaps some rejecting it.

The model also has the consequence that the same person would also have the same degree of belief in flower *'f* is yellow'. But borderline cases are meant to be a bit tricky, and I think it's implausible that an agent will always have the same very precise tendency to act as if *f* is yellow. That degrees of belief come out so precisely on this model is a part of the problem, and looks quite strange. If every agent has the exact same degrees of belief about a vague proposition, they would all be equally confident in their judgement of just how true it is, just *how* yellow flower *f* is. This isn't a plausible way to handle borderline cases. Throughout the thesis I have taken them to ones that are difficult to judge in some sense, and that doesn't seem well modelled by a theory that says every agent will have the same precise degree of belief about them.

Whilst I don't object to the use of objective truth values in Smith's model in general, that the overall model gives the result that agents will have the same precise degrees of belief (to the same propositions under the same circumstances) makes it implausible, and for that reason I will prefer a theory of belief to be now be considered, offered by MacFarlane. According to this theory, beliefs about vague propositions are not precise, but rather indeterminate. This is a more attractive way to treat borderline cases, and allows for differences between agent's attitudes, whilst keeping belief and truth values connected.

3. MacFarlane and Partial Belief

MacFarlane (2010) defends a model which is similar to the one just discussed. It satisfies what I am looking for from a cognitive theory for degree semantics here, as it takes rational belief for the degree theorist to be a kind of partial belief, and gives an account of how these partial beliefs interact with probabilistic partial belief. MacFarlane also endorses that there are two kind of partial beliefs, following Schiffer (2000). One is *standard partial belief* (SPB). These are the beliefs standardly associated with probabilistic belief, beliefs we have of which we are less than certain due to our epistemic circumstances. These are the sort we have called *uncertainty* so far. The other is *vagueness-related partial belief.* These are the sort you can have even in ideal epistemic circumstances, when judging a borderline case.

However, MacFarlane makes a diversion from the sort of path Smith takes at the outset. As we've seen, in Smith's model, agents degree of belief in a vague proposition is represented by a single precise value, matching either its degree of truth (or perceived degree of truth). MacFarlane thinks a better approach is to take it that agents are unsure of what exact degree

of truth a vague proposition has, and thus allow our model of belief to reflect that. Accordingly, he models belief with a *probability distribution* (2010, in Dietz and Moruzzi 2010: 449). This is most clearly represented on a graph; with credence on the vertical axis and degrees of truth on the horizontal axis, you get a curve showing the probability a subject invests in each degree of truth. For some vague proposition, 'Tom is bald', and for some agent, we could generate a curve, the height of which over each degree of truth represents the probability the agent invests in 'Tom is bald' having that degree of truth. For a clear case of not baldness, the graph would show credence to be at 0 for all degrees of truth, except for degree 1. For a clear case of baldness, the graph would show belief to be at zero for all degrees of truth except 0. For a borderline case of baldness, an agent may well be aware that the degree of truth of that proposition is neither 1 or 0, but falls somewhere in between. However, they may not be able to tell exactly what the degree of truth is:

> We will never be in a position to know who is the shortest man who satisfies 'tall man' to degree 1, and we will have no good basis for taking the proposition that Jim is tall to be true to degree 0.653 rather than 0.649. We may be confident that he satisfies 'tall' to some indeterminate degree, and perhaps we'd bet on 0.6 over 0.5, but there will remain some uncertainty. (MacFarlane 2010, in Dietz and Moruzzi 2010: 449)

Thus, for any borderline proposition, we are already in a situation of uncertainty, uncertainty about the exact degree of truth that borderline case takes, and a probability distribution models that attitude, representing the credence invested in each possible degree of truth.

Seeing Tom, with 350 hairs on his head, our attitude towards 'Tom is bald' can be represented by a probability distribution, reflecting how likely we think it is that 'Tom is bald' is true to each possible degree of truth. If Tom is a typical borderline case, and actually bald to degree 0.5, it might be that our credence is quite high around the 0.475-0.525 range, and tails off around 0.4 and 0.6, with very low credences in the ranges from 0 to 0.3 and 0.7 to 1. What do we do when we aren't sure of how many hairs Tom has, and thus are affected by uncertainty also? Say last time I saw Tom he had 350 hairs, but I haven't seen him for a month, and I know that he is currently losing hair regularly. I am uncertain of how many hairs he has now. I think he will have lost *some,* so I think it is unlikely he has 350 still, and I doubt that he will have lost all or most of them. MacFarlane suggests that this attitude can also be modelled by a probability distribution; a curve representing for each possible number of hairs, how likely it is that Tom has that particular number of hairs on his head. In this case it might start very low in the 0-100 region, rise and peak around 200, and drop off to very low by 300-350 again. Given that uncertainty can be modelled by probability distributions, and

our attitudes to particular vague propositions can also be modelled by probability distributions, the combination of the two attitudes results in a simple application of probability theory (2010, in Dietz and Moruzzi 2010: 449-454). Our probability distribution for how many hairs we think Tom has includes assigning a probability to each individual possible number of hairs (in a given range). For each possible number of hairs that Tom could have, there is an individual probability distribution, representing the likelihood we take of 'Tom is bald' having each possible degree of truth given that he has that particular number of hairs. Thus, the probability distribution over how many hairs we think it is likely that Tom has, in fact gives a probability to different probability distributions being the one we have towards 'Tom is bald'. Thus, we can combine this into one single probability distribution with the following formula:

Pr([Tom is bald] = x) =

$$\sum_{0 \leq n \leq 350} Pr(Tom\ has\ n\ hairs) \cdot \Pr(Tom\ is\ bald) = x\ given\ that\ (Tom\ has\ n\ hairs)$$

To get the probability distribution that represents our attitude towards 'Tom is bald' given the likelihood we assign to Tom having each particular number of hairs, we sum the product of the probability of Tom having some number of hairs and probability of 'Tom is bald' being true to some degree given Tom has that number of hairs, for each degree of truth, for each possible number of hairs. The result is a curve over degrees of truth, showing the likelihood of each degree of truth given the probability we assign to each possible number of hairs.

MacFarlane claims that this method gives a straightforward solution to the problem of combining two different sorts of partial belief. If we treat attitudes to both vagueness and uncertainty as probability distributions, then we can use probability theory to combine the two quite easily.

So here we have the outline of an alternative theory of rational belief for the degree theorist. Belief in a vague proposition is again a partial belief, but is represented by a probability distribution over possible degrees of truth, rather than a single precise value. This kind of partial belief can be easily combined with probabilistic credences, through the combination of probability distributions in the way just described. MacFarlane doesn't give an account of exactly it *means* for an agent to belief a proposition to degree 0.6 or 0.5, but we can reasonably continue with the account used by Smith, that degrees of belief represent a sort of *strength of tendency to act.*

3i. Features of the model

There are some interesting consequences that come with using probability distributions instead of expected values as a model of belief. In the case of Smith's model, belief was always a precise value, matching expected truth value. One of the consequences of this is that rational belief follows the same patterns as truth values in the degree semantics. However, with probability distributions there are some subtle differences.

For instance, MacFarlane argues that this model comes out with the right valuation for *complex* sentences involving partially believed vague propositions, which is actually an improvement on the treatment offered by probability theory from the classicist, *and* by degree theorists using Lukasiewicz style valuations (2010, in Dietz and Moruzzi 2010: 452-3). According to the semantics, conjunction follows the *min* rule: the truth value of P&Q equals the value of P or Q, whichever is lower. This is the same pattern belief would follow on Smith's model. In MacFarlane's case, our attitude to P&Q will be a probability distribution. The degree of truth of P&Q is still equivalent to truth value of whichever proposition has the lesser truth value, and the probability of P and Q having truth values *d1* and *d2* respectively is equivalent to the product of the probability that P has truth value *d1* and the probability that Q has truth value *d2*. To get the probability that P&Q has truth value *x,* we sum the probability of each possible combination of truth values for P and Q when the minimal value equals *x*.

The result of this is that our degree of belief in a complex sentence involving vague propositions won't equal the degree of belief in individual propositions; instead it will be slightly less than any of the particular propositions. It does however depend on the probability distributions for the conjuncts. The more uncertainty about the truth values of the conjuncts, the less predicted confidence in the conjunction: the further the probability distribution of the conjunction will be shifted towards lower degrees of truth. If one were *certain* of the truth values of P and Q, then the above model would predict that one would be certain that the truth value of the conjunction of them equalled the truth value of which ever was the lower of the two. However, MacFarlane think this isn't possible; one is never certain of the truth value of vague propositions (which are borderline cases at least). Thus we will always believe the degree of truth of a conjunction of vague propositions to be slightly lower than we believe the degree of truth of any one conjunct to be.

MacFarlane suggests that this is a desirable result, but doesn't give a full explanation of why. It is worth noting that earlier in the same paper, MacFarlane argues that degree semantics are a better model of vague discourse that classical semantic, and an important aspect of this argument was that the standard degree semantics for conjunction hold (i.e. that degree of belief of a conjunction would equal the lowest degree of belief invested in either conjunct) (2010, in Dietz and Moruzzi 2010: 445). The general idea was that an epistemicist making use of *classical* logic and semantics is committed to saying any partial belief in a vague proposition is the product of uncertainty; after all, every proposition is either true or false, if we are uncertain of which it is because we are ignorant of the facts. If this is right, then where an agent believes vague propositions P and Q to degree 0.5, they believe the conjunction P&Q to degree 0.25, as those beliefs are probabilistic, on account of being caused by uncertainty. MacFarlane claims this isn't the right result, rather, the right degree of belief in the conjunction is 0.5, as matches its truth value according to the Lukasiewicz semantics, and thus degree semantics is a better model of vague discourse. Here it mattered that degree of belief in a conjunction did equal the lower of the two degrees invested in each conjunct, but MacFarlane now claims it should be slightly less than that. Of course, how much the degree of belief does drop depends on how much uncertainty there is, and where the agent is certain there is no drop at all; perhaps MacFarlane meant for us to understand the argument to be a case where there were ideal epistemic circumstances (even though no such cases are really possible in his opinion). But this then seems to beg the question against the epistemicist, as in their opinion an agent would never have a partial belief in ideal epistemic circumstances. It is worth noting this interesting feature of using probability distributions to model belief, but MacFarlane hasn't done enough to motivate why it is a benefit here.

Using probability distributions can also generate some resources for dealing with concerns about truth functionality. A common complaint levelled against truth functional degree semantics is that it gets the wrong truth values for some complex sentences. This includes the possibility of contradictions that aren't definitely false, and instances of breaking the law of identity that aren't definitely false. For example, if the truth value of a conjunction is equivalent to the truth value of the conjunct with the lower truth value, then the truth value of (P&¬P) is 0.5 where the truth value of P is 0.5. MacFarlane's model doesn't offer a novel approach to those concerns, but he argues it does for a different concern about the truth functional conjunction (2010, in Dietz and Moruzzi 2010: 459-60). Consider two borderline cases of tallness, Jim, who is tall to degree 0.5, and Tim, who is tall to degree 0.45. Following

this slight difference in truth value, Jim is slightly taller than Tim. On truth functional degree semantics, the truth value of 'Tim is tall and Jim is tall' is 0.45. However the truth value of 'Tim is tall and Jim is *not* tall' is also 0.45. But if Jim is taller than Tim, then surely if ever Tim is tall, Jim has to be also; there is no chance of Tim being tall while Jim is not. This is considered a conceptual fact about 'tallness', that anyone taller than someone who is tall is necessary tall. So it is argued that truth functional degree semantics get the wrong result here.

MacFarlane offers some relatively standard responses to these sorts of concerns, attempting to make the results appear less counter-intuitive. After all 'Tim is tall and Jim is not tall' is true to degree 0.45, so it more false than it is true. And surely we shouldn't expect it to be true to degree 0, as then presumably 'Tim is not tall or Jim is tall' would be true to degree 1, but that doesn't look appealing if neither is assertible. MacFarlane goes on to say that nevertheless, there does seem to be an important difference between 'Tim is tall and Jim is tall' and 'Tim is tall and Jim is not tall'.

MacFarlane suggests that his model can offer an explanation. If we model attitudes towards the propositions not as precise degrees of belief, but as probability distributions, then we can find a difference between the two conjunctions. Instead of having degree of belief 0.5 in 'Jim is tall', and 0.45 in 'Tim is tall', each is a probability distribution. MacFarlane models both as a normal distribution, centred on 0.5 and 0.45 respectively. The probability distribution of 'Tim is tall and Jim is not tall' shows that it is taken that there is no chance of it being truer than degree 0.5, while distribution for 'Tim is tall and Jim is tall' has a tail that reaches to degree 0.8. The model predicts that we think the latter could be true to a degree which there is no chance of for the former. MacFarlane contends this can explain why we have different attitudes to the different conjunctions, even though truth functional semantics assign them the same degree of truth. Of course, we might worry that this doesn't deal with the main force of the objection. One might reasonably complain that we think there is no chance of a contradiction being true to any degree other than 0, no allow some belief that it could be true up to degree 0.8, even if never fully true. In this sense I don't think the probability distributions give the means to fully avoid the worry for degree semantic, but it is interesting to note the none the less.

MacFarlane uses a similar resource to defend from other worries about degree semantics (2010: 462). Edgington argues that if 'Tom is bald' is true to degree 0.5, we might expect the truth value of 'Tom is bald and tall' to vary as we vary the degree of truth of 'Tom is tall'

between 0 and 1. The upshot of this is that the conjunction should have a higher degree of truth when 'Tom is tall' is true to degree 0.7 than when true to degree 0.5. But on the truth functional semantics, 'Tom is bald and tall' would have the same degree of truth in both of those situations, degree 0.5. MacFarlane argues away the intuition that they should have different degrees of truth, but suggests that it might be right that we have different attitudes to 'Tom is bald and tall' when 'Tom is tall' is true to different degrees above 0.5, even though the truth value of the conjunction doesn't alter, and that this can be captured by probability distributions. Assuming normal distributions for 'Tom is bald' and 'Tom is tall', if we fix the distribution for the former as centred over 0.5; we get different distributions for the conjunction of the two when the latter is centred over 0.5 and 0.7. When the distribution for 'Tom is tall' is centred over 0.7, the distribution for the conjunction is centred just below 0.5, whereas it is closer to 0.4 when the distribution for 'Tom is tall' is centred over 0.5. As before the effect is due to the uncertainty over the values of the conjuncts, and with more uncertainty we would see a greater shift of the distribution for the conjunction to lower degrees of truth.

These are some interesting consequences of using probability distributions to model belief rather than a single expected value, in the way that Smith does. In Smith's model, belief follows the same pattern as truth in the degree semantics, and thus doesn't offer any new ways to reply to these standard concerns for degree theoretic approaches to vagueness. That is one consideration in the comparison between the two. But how tenable is this approach to the theory of mind for degree semantics?

3ii. Criticising MacFarlane's model

Probability distributions are introduced very quickly in MacFarlane's paper, and little is said about how we should think about them as representing attitudes, and how different distributions interact with things like action, decision making and reasoning. MacFarlane leaves many questions unanswered here. Using probability distributions may give a convenient solution when it comes to combining different sorts of partial belief, and gives an interesting alternative to a theory like Smith's, but we are far from understanding what the consequences of using probability distributions are.

One concern is that we don't know what it really means for a probability distribution to represent an attitude to a proposition. Is it a kind of complex belief? Or meant to represent

some sort of indeterminacy in a simple partial belief? A related worry is that it's not clear how probability distributions line up with action and decision making. Smith's model can give more straightforward answers in this case, and I worry that there is a pressure for MacFarlane to suggest that actions aren't really guided by the probability distributions, but the *expected values* we can get from them. If this is the case then it seems to be a departure from MacFarlane's view, and would leave us wondering why bother with the distributions in the first place.

So how we should understand the attitudes represented by a probability distribution? Consider an agent who has a partial belief in a vague proposition P, which is represented by a probability distribution which is centred at 0.5, with tails starting at 0.3 and reaching 0.7. What does this say about the agent's attitude towards P? There are two main sorts of possibility:

1. The agent has a complex attitude towards P; she thinks there is some possibility that P is true to degree 0.3, and some possibility it is true to any degree up to 0.7, but that it's most likely true to a degree around 0.45-0.55
2. The agent has a fuzzy sort of attitude to P; the agent doesn't believe P to be true to any of the degrees between 0.3 and 0.7, rather it is indeterminate to what degree she believes P is true. Here the agent could be disposed to act as if they believe the proposition is true to any of the relevant degrees of truth, and the probability curve represents the likelihood each degree of truth has

MacFarlane isn't very explicit about what sort of direction he intended to go in, but I take it that he has something like option 1 in mind. He is discussing a position on which we are necessarily uncertain of the truth values of indeterminate sentences, and thinks the distributions capture that well; we are uncertain of which degree of truth P has, and this uncertainty is modelled by a distribution neatly. It's a view where P will always have a determinate truth value, and the ideal degree of belief to invest in P is the one which matches its degree of truth, but agents are unsure exactly where it falls. This is also the interpretation that fits most clearly with the combination of partial beliefs from both vagueness and uncertainty; in both case the agent invests varying levels of credence in different possibilities. This is slightly to different to option 2, where agents have a something like a *mushy* degree of belief; it's imprecise where the degree of belief falls. It may well be aimed at the degree of

truth of the proposition, but there isn't a determinate attitude; opposed to option 1 where there is determinate credences invested in each possible degree of truth

Option 1 is complicated, as it isn't clear how it should connect with action as a probability distribution. If this is the option MacFarlane has in mind, I'm worried that it looks most promising to reduce the curve to an expected value, threatening a collapse to a model more like Smith's. I'll argue option 2 can give us a way to distance the probability distributions further from Smith's model, and thus that this is the preferable way to interpret MacFarlane's model.

3iii. Probability distributions and decision making

So how should we think probability distributions connect with action, given an interpretation like option 1 above, where the agent has a complex attitude towards a vague proposition? Here we are taking degrees of belief to link to strength of tendency to act, but with a complex attitude, the agent doesn't have any particular degree of belief, but instead has many partial beliefs that a proposition has different possible partial degrees of truth. How will this translate to action?

Say that I'm going on holiday and I want to take all of the *short* novels I have with me. How likely am I to pick *Ulysses*? Extremely unlikely. How likely am I to pick *Breakfast at Tiffany's*? Very likely. Here I will have a full belief that *Ulysses* is not short, and a full belief that *Breakfast at Tiffany's* is short. A probability distribution isn't really necessary in these cases, but it would show that I take there to be no chance of '*Ulysses* is short' having any degree of truth other than 0. Similarly, a probability distribution for '*Breakfast at Tiffany's* is short' would show that I think there is no chance of it having any degree of truth but 1.

But things aren't so easy with borderline cases, with a more complicated probability distribution. How likely am I to take *1984*, given that I take it to be a borderline case of a short novel? We would like to be able to say that the higher my degree of belief that *1984* is short, the more likely I am to take it. This answer isn't problematic for a model where our degree of belief is represented as a precise value. On Smith's model, we can say that the higher the degree of belief in P, the more likely we are to act as if P is true. Thus I am more likely to take book with me I believe to be short to degree 0.8, than one I believe to be short to degree 0.7.

On MacFarlane's model, this approach is not so readily available, as attitudes aren't modelled as single degree of belief. Given option 1, understanding the probability distribution to represent a complex belief varying credences over a range of degrees of degrees of truth, it is unclear just how taking the truth value of a proposition to fall in this certain range interacts with action. Say for example, I have space in my bag for one more item, and have to choose between two remaining novels, which do I pick? The simplest answer is the one which I believe to be short to the greater degree, which is the one that Smith's model could produce. But this isn't an option for MacFarlane, as on his model we don't believe either novel is short to a particular degree, but just think both are to a degree in a particular range. But then how do we decide between two novels? I am uncertain of the degree of shortness of each book.

In some cases we might have answer. Say that I take the truth value of '*1984* is short' to be between 0.6 and 0.8, and I take the truth value of '*Frankenstein* is short' to be between 0.3 and 0.5. In this instance I believe it to be impossible for the degree of shortness of *1984* to be less than the degree of shortness of *Frankenstein,* thus here I can easily make decision based on my probability distributions, I pack *1984.* But what about if I were take the truth value of *Frankenstein* to be between 0.5 and 0.7? In this case I'm not sure of the degree of shortness for each book, and think that either *could* be the shorter novel.

Perhaps in such cases I just have to make an *arbitrary* choice. We just don't believe either book is shorter, we are necessarily uncertain; we have to make a decision as we do in other cases where we are uncertain. This may not be so bad. After all, on a model where belief is represented by a precise degree we still may have to make arbitrary choices. If for example I had the same degree of belief that *Frankenstein* was short, as I did that *1984* was short, and being short was the only relevant consideration when deciding which novel to pack. In this case I may well have to make an arbitrary decision about which book I choose. Perhaps on MacFarlane's model, where we are ever uncertain of which novel is short to a greater degree, we have to simply make an arbitrary choice.

This doesn't seem quite right though. You can understand that where your attitude towards two different novels is identical an arbitrary choice may be necessary, but where it's not, it seems that an arbitrary decision shouldn't be necessary, especially given an alternative. The most natural thought in the case above, where I take the degree of shortness of *1984* to be between 0.6 and 0.8, and of *Frankenstein* to be between 0.5 and 0.7 (assume for now both are normal distributions), is to take *1984,* as I think it most likely that the degree of shortness of

1984 is greater than that of *Frankenstein*. The intuition to view it this way is what speaks most clearly, but to do so requires a what appears to be a departure from MacFarlane's view of probability distributions, to the use of *expected values*.

To settle the dispute in this way is to suggest that we take the book which we think is more likely to be shorter. In the case currently under inspection I take the degree of shortness of *1984* to be between 0.6 and 0.8, and of *Frankenstein* to be between 0.5 and 0.7, and both are normal distributions. It is natural to think I would choose *1984*, but I think it likely that its degree of shortness is greater than that of *Frankenstein*. But here I am using the *expected* value for each; an average of the possible degrees of shortness for each book, weighted by the probability of each degree. The expected degree of shortness of *1984* would be 0.7, whereas it would only be 0.6 for *Frankenstein*. This would give a straight forward way of giving a solution to the concern I have, but it seems to be an important step away from MacFarlane's view, and towards Smith's. If we need to use expected values to be able to determine how a probability distribution relates to decision making, then we might wonder why we should bother with the distributions in the first place? After all, Smith allows there to be a number of different possibilities in play, just as there are in a probability distribution, but collapses them to a single value. It's not clear what the benefit of MacFarlane's model is if we need to take the step Smith also does to fully make sense of it.

For instance, take the case in the florist, where you hide 4 flowers behind a curtain, which are yellow to degree 0.2, 0.4, 0.6, and 0.8, and hold one obscured from view. What is my attitude towards 'you are holding a yellow flower'? For MacFarlane, my attitude can be represented by a probability distribution, in this case most likely a series of four smaller normal distributions, with means over 0.2, 0.4, 0.6, and 0.8. For Smith, we can represent my attitude by the expected value, in this case 0.5. But as there is uncertainty at play, we could also go back a step and also create probability distribution. In this case, it would likely be 4 sharp spikes, over 0.2, 0.4, 0.6, and 0.8. If we have to use an expected value to make sense of the way MacFarlane's probability distribution affects behaviour, then what is the substantive difference between these two models? I'm not sure there is one. Given the current example, they disagree about where there is uncertainty, which is shown in the difference in shape of the distributions; MacFarlane has smooth curves opposed to spikes, as he expects uncertainty around what the degree of truth is for each flower, whereas Smith expects agents to select a particular value. This may mean they also disagree on what the expected value is in each case

(depending on what MacFarlane's curve actually is in each case), but again, we may start to wonder in what ways MacFarlane's theory is substantively different from Smith's.

MacFarlane does mention the possibility of a model which uses an expected value briefly (though only in a footnote). He writes:

> While it might be useful for some purposes to measure beliefs by the expected value of degree of truth, it seems rash to suppose that all the interesting quantitative differences between partial belief states can be boiled down to this one number. For example, compare (a) a belief that assigns equal credence to each degree of truth (a flat line on our graphs), (b) a belief stat that assigns near certainty to degree 0.5 (a sharp spike), and (c) a belief start in which credence clusters around two points, degree 0.2 and degree 0.8 (two humps with a dip in the middle). In all three cases the expected value of degree of truth will 0.5, but these belief states can be expected to have different effects on behaviour and inference. (2010, in Dietz and Moruzzi 2010: 453, *fn* 13)

Here MacFarlane suggests that he thinks in certain situations it would be useful to think about attitudes as expected values, and so perhaps would be ok with it in a case like my novel dilemma above. But he also gives a reason to keep probability distributions: attitudes are complicated and can't always be summed up in one number; we can expect different behaviour even when people's attitudes would be represented by the same expected value.

But this is unconvincing, particularly when MacFarlane hasn't told us anything about how we can expect probability distributions of any shape to interact with behaviour. It might seem initially plausible that say a sharp spike and a flat line would lead to different kinds of behaviour, but it's not entirely clear how. Say again I am choosing a final book to take on holiday, and I have to choose between *1984*, and a mystery book hidden in a box. I am nearly certain that '*1984* is short' is true degree 0.5, thus have an attitude towards that proposition represented by a sharp spike at 0.5. I have no idea what the mystery book is, so my belief that it is short is represented by a flat line across all possible degrees of truth. In what ways in each likely to affect my behaviour? From the distributions alone, it's not clear which book I am most likely to pick. In this case it may well come down to different factors, such as how risk averse I am; where I desire a short book, I may or may not prefer to gamble that the mystery book will be shorter than *1984*. But for things to be decided by risk aversion we are already accepting that we cannot decide just on our beliefs alone, given that they have the same expected value. If I were almost certain that *1984* were short to degree 0.6, it would likely be the case that I'd choose that over the mystery book, as it is unlikely it will be shorter than *1984*. But things like risk aversion can be just as relevant in Smith's theory where we focus on the expected value. For Smith, as we cannot choose between *1984* and the mystery

book based on our strength of tendency to act alone, risk aversion could come in as a deciding factor. So again, it's not clear what MacFarlane's theory has gained here.

Furthermore, here we're primarily interested in attitudes towards vague propositions. When faced with a borderline case, what probability distributions are likely to come about in an agent? It is surely not two humps (option c in the quote). For MacFarlane we are unsure of where *exactly* the degree of truth of a borderline case falls. It is unlikely we will think, for example, it is most likely either 0.2 or 0.8, but unlikely to be anything in between. A flat line is surely not a likely attitude when faced with a borderline case. If I see a borderline case, it's not that I have no idea which degree of truth is likely, and think each has an equal chance. Agents can tell if a borderline case is closer to truth or falsity, or somewhere towards the middle. A sharp spike is also not a likely attitude, at least on MacFarlane's theory, as he takes it that agents will always have some uncertainty over the degree of truth of a vague proposition. I think we should not expect most attitudes towards borderline cases to stray particularly far from normal distributions. It appears, following MacFarlane's quote above and what is intuitive, that for a borderline case, an agent will have a range where they feel the truth value is most likely to be. This would be the top a bell curve, and as the truth values move further from that range on either side, so would the credence invested by the agent (at what rate I am not sure).

But if there is a pretty standard shape when it comes to probability distributions representing attitudes to vague propositions, then MacFarlane's argument above is less compelling. While it is true that probability distributions with significantly different shapes but the same expected value may well lead to different behaviour and inference, it seems unlikely that such different shapes would be induced by borderline cases. And if expected values do the work when it comes to decision making with probability distributions that are standard distributions, then MacFarlane's reason here is not a good one for preferring probability distributions to expected values.

For these reasons, I think that interpreting probability distributions in the way we've described as *option 1,* a complex attitude representing different invested levels of credence in possible degrees of truth, isn't the best way to go. The best way to understand the way the distributions interact with action in that case is with expected values, and then it seems MacFarlane isn't offering a genuine alternative to Smith.

3iv. Probability distributions and indeterminate belief

What about the prospects of *option 2*? On this view, the probability distributions represent that the attitude is indeterminate between a range of values, but that there is a difference in weighting between them. I think that here can find a connection to action that keeps the theory a genuinely distinct from Smith.

Take again the case where I'm looking to take my short books on holiday with me. When I'm considering whether to take each of the novels I own, I will have an attitude represented by a probability distribution. On the current suggestion these show an indeterminacy in belief, between a range of values. How does this play out in action? One way to interpret it is to take it that an agent is disposed to act as if they believe the relevant proposition to any of the degrees covered by probability curve, and that the variance in the credences shown by the curve show how likely each it is that the agents acts in accordance with any single value. For example, consider a normal distribution centred over 0.5, with tails reaching 0.3 and 0.7. For any degree of truth, the curve shows the likelihood of acting in accordance with that degree of belief. For instance, say that the peak of the curve over 0.5, reaches 0.3 probability, and thus I will act as if I believe the relevant proposition to degree 0.5 30% of the time. This being a normal distribution, there is a good chance I will act as I believe it slightly more or slightly less. The values towards the tails of the curve are much less likely; it may be a percentage close to 0 that I act as if I believe the proposition to degree 0.3 or 0.7. Where the curve reaches 0 (here degrees of truth 0-0.29, 0.71-1), there is no chance of acting in accordance with those degrees of belief. How likely am I to take *1984*, given that I take it to be a borderline case of a short novel? In the case of *option 1*, we could find no better answer than to revert to the expect value of the probability distribution to determine the strength of tendency to act. But in this case, we can say instead that behaviour will vary depending on the probability distribution. If the probability distribution representing my attitude to *'1984* is short' is like the one just described, a normal distribution with a mean of 0.5 and tails reaching 0.3 and 0.7, then I am disposed to act with a strength of tendency corresponding to any of those degrees, but most likely will act with a degree of belief around 0.5.

This comes with some initial benefits, primarily that it allows for some variation when faced with the same borderline case. It doesn't predict that an agent will always treat a borderline case in the same way, but may act with different strengths of tendency on different occasions, which wasn't so clearly possible with Smith's model. Smith's might allow that an agent will

make a different decision on different occasions (where we are associating strength of tendency to act with a *frequency* of acting in some way), but otherwise suggests that agents will always have the same strength of tendency to act. With the current interpretation of probability distributions, an agent needn't always have the same strength of tendency to act, it can vary from occasion to occasion, as the agent has a sort of indeterminate attitude towards it. Importantly, different agents will also not have the same precise attitudes, but rather indeterminate ones, which will likely result in treating borderline cases in different ways (with different strengths of tendency as each other).

Things in the comparison case still are bit unclear. When I have to choose a book between *Frankenstein*, which I take to be short to a degree between 0.6 and 0.8, and *1984*, which I take to be short to a degree between 0.5 and 0.7, which do I pick? Again there is some pressure to go towards expected values again, but I don't think that does full justice to the suggestion that these distributions represent a kind of *indeterminate* attitude. In this case I think either could be shorter, and I have dispositions such that I could act as if either is the shorter model. Perhaps in these cases it comes down to however my behaviour happens to play out on the occasion. The distributions show that I'm most likely to take *Frankenstein*, as I'm disposed to act more strongly as though it is short than *1984*. But on a day where I happen to act more weakly that *Frankenstein* is short and more strongly as though *1984* is short, perhaps I would take *1984*.

The biggest issue with this interpretation of probability distributions is that it interferes with the straightforward combination of partial beliefs from vagueness with partial beliefs from uncertainty. For instance, consider the case where you are obscuring four borderline yellow flowers, one of which you will be holding, and my attitude towards 'you are holding a yellow flower' is represented by a probability distribution with a four separate normal distributions centred over 0.2, 0.4, 0.6, and 0.8 respectively. In this case I'm disposed to act as if 'you are holding a yellow flower' over a large range degrees, perhaps from 0.1 to 0.9. But that doesn't look so plausible. As I think there's an equal chance of you holding any of the flowers, I'm equally disposed to act in accordance with degree 0.2, 0.4, 0.6 and 0.8, but that looks plainly wrong. In these cases it shouldn't be that the combined probability distributions represent a single indeterminate attitude. It should represent the indeterminate beliefs I have with respect to the four different possibilities.

So what attitude do I take towards 'you are holding a yellow flower' in this instance? In an attempt to stay true to the idea that this is a combination of indeterminate belief, instead of reverting to the expected value of the probability distribution, we can allow that the expected value will vary with the respective dispositions for each possibility. Thus no particular single value can represent my attitude to the proposition, as would follow from *option 1*, or Smith's model. Just as my attitude can vary on occasion for any particular vague proposition, the expected value of a vague proposition based on the combination of more than one probability distribution can also vary on occasion. In this current case, the expected value will be based on whatever individual value is determined for each possibility. I'm disposed to have a degree of belief around 0.2 for flower 1, around 0.4 for flower 2, 0.6 for flower 3, and around 0.8 for flower 4, but these can vary from occasion to occasion on the current proposal. The expected value will be the product of the relevant values on the occasion divided by four (somewhere around 0.5 in this case).

This is the natural move given the current understanding of probability distributions as indeterminate attitudes, and gives us a genuine alternative to Smith's model, unlike understanding probability distributions as a kind of complex belief. Here attitudes towards borderline cases are a sort of indeterminate belief, represented by a probability distribution. Agents are disposed to a range of degrees of belief in vague propositions, with the probability curve representing the likelihood of acting in accordance with any one. We can combine probabilistic beliefs with partial beliefs induced by vagueness still, but may have to resort to expected values in some cases; but these expected values are indeterminate in a sense also, as they vary just as attitudes to each possibility vary.

Smith's model contrasts this. Belief is always determined by a particular degree of belief and strength of tendency to act, which remains constant over different occasions. Whilst both models have given a possible way to expand on theory of mind for degree theoretic approaches to vagueness, it is clearly MacFarlane's theory that is preferable. A model which takes agents to have precise degrees of belief matching the degree of truth (where there is no uncertainty), and always the same degrees of belief as each other, is not a plausible one. MacFarlane's theory gives us the result that degree of belief in vague propositions isn't so precise, which is a more natural way to handle borderline cases, and allows the more plausible result that agents wont all have the same precise strength of tendency to act in each case.

4. Conclusion

I've examined two possible ways of developing a theory of mind for degree semantics, and have argued that MacFarlane's theory is the better option. It involves partial belief, which can be connected with action through strength of tendency to act, however doesn't have the implausible consequences Smith's theory did.

These last two chapters have been focussed on investigating what theories of mind are compatible with two popular theories of vagueness. I argued that supervaluationism should be associated with a suspension of belief view, and we now have MacFarlane's model to associate with degree semantics. In the following final chapter, I will take these results back to a more direct discussion one the experimental approach to vagueness, considering that the upshot of accepting these theories of mind is when examining empirical work.

As argued in chapter 3, past studies haven't developed a satisfactory link between theories of vagueness and cognitive theories. I argued for such a link in chapter 4, and have now examined theories of mind for supervaluationism and degree theories. I will return to those past results, and review them with these new cognitive models in hand; we can approach them from a new position of theoretical strength, having dedicated time to the connection between theories of vagueness and cognitive theories.

Chapter 7 Conclusions: Taking Stock and Looking Forward

In this final chapter I want to bring some of the threads of argument in this thesis back together. I'll recap some of my main conclusions, and see where they leave us with regards to vagueness and the experimental approach. Part of what I've argued is that the different logics and semantics suggested in theories of vagueness come with cognitive commitments, and went on to examine what these commitments look like for supervaluationism and degree theories. Here I will discuss in more detail how we should expect these cognitive theories to play out in the experimental setting. I will also return to the experiments I discussed, where the theorists failed to develop a substantial connection between theories of vagueness and theories of mind, to review the results in light of my suggestions. Finally, I'll consider the future of the experimental approach, and how it might be best to continue with the foundations I suggest here.

1. Taking Stock

Here is a quick recap of how the argument in the thesis has developed and the key conclusions in each case:

The experimental approach can inform theorising about vagueness

The most general aim of this thesis has been to defend the experimental approach to theorising about vagueness, and develop the right way of going about it. I argued that the experimental approach can make a theoretically informative contribution, because the goal of philosophical inquiry about vagueness was a *descriptive* theory, which captures the logic and semantics of real language use. I suggest that a theory which was primarily prescriptive would be largely uninteresting. This is the sort of project I attribute to Frege (1879, 1903), or the field of conceptual ethics (Burgess and Plunkett 2013), and it entails a redefinition of our concepts. This might seem a worthwhile pursuit for some, as it was for Frege, but I think what's really interesting about the problem of vagueness is speakers successful use of vague language despite its apparent paradoxical nature. It isn't necessary to abandon natural language and concepts, as it works just fine, the project is to explain how it works, given that it seems this should be problematic. This is a descriptive question, and I take it that this is the question that most philosophers are actually engaged with. Philosophers explicitly aim to

keep their theories connected to vague language use and standard intuitions about vague concepts, and criticise theories when they fail to do so. As philosophers are trying to give theories that represent actual language use in this way, I argue there is space for empirical work to provide more information about the language use theories are trying to capture. As noted, there is a sense that the descriptive theories have a normative element, in that they give a theory of successful language use. But it's important to note that what dictates *successful* language use is the usage of normal speaker. It is possible to be unsuccessful with the use of any term in natural language, to misapply 'table' for example, and what determines the correct usage of 'table' is the linguistic conventions of the community. The same thought is behind the suggestion here. There is a correct and incorrect usage of vague language, and theories of vagueness are part of capturing the correct usage. But what determines correct usage is linguistic conventions, how speakers actually apply the language, keeping the theories tied to real life language use.

Experiments must be done in the right way

As the philosophical investigation into vagueness is a descriptive project, I argue that there is the potential for experiments to play an informative role. This a priori argument only shows that studies may be useful, but the reality of finding an important result is very difficult. In chapter 3, I gave three examples that I argue fail to establish their conclusions (Bonini 1999, Alxatib and Pelletier 2012, Egré et al 2013). We saw some common themes which made the studies unsatisfactory. A good study must have appropriate statistical methods, which most did not. Without this, we cannot truly assess the strength of findings in the data. It isn't enough to know that *x*% of participants report *y*, we need a statistical test to show how likely it is that result was down to chance. Studies also failed to have a well-established account of what different theories of vagueness would predict. To really make progress with the experimental approach we need a good theory of what we should expect to see in empirical studies from the perspective of popular theories of vagueness, and why. There is a lack of a well thought out theory of the cognitive consequences of theories of vagueness and a lack of argument about whether it is reasonable to suggest there are such consequences, in the contemporary literature. To get the experimental approach going, we need a good argument that we can link different responses in studies with theories of vagueness, as well as an account of which responses connect with which theories. This side of the theoretical background hasn't been clearly enough developed in recent studies. We saw for instance, connecting a denial of the law of excluded middle being associated with a supervaluationist-

type response to borderline cases, despite the fact that law of excluded middle is tautologous in supervaluationist logic. To really be able to make claim to an interesting finding in an experiment, we need to fill in this side of the story: can we suggest that theories of vagueness make particular predictions about the results of experiments, and what are these predictions?

Semantic theories have commitments in the theory of mind

The second part of the thesis aimed to develop a theoretical basis to answer those questions just posed. It's not enough just to take it that theories of vagueness are descriptive, to be able to connect a theory to particular experimental results we need to know what they are committed to saying about language use. I suggest we can make this connection because logico-semantic theories some restrictions on theory of mind, contra to the views of some philosophers (e.g. Schiffer 2000). This is because 'truth' and 'logic' come with cognitive commitments. Belief aims at truth, and logic gives the details of necessary truth preservation. Thus the details of truth and logic in a theory restrain what that theory says about belief. If these connections are denied we can start to worry about whether some different theories are distinct in a substantive way, as suggested by Williams (2014). If two theories claim to give a different notion of truth and a different logic, but have the same theory of mind, i.e. the same rational attitude to all the same propositions, we can worry that the apparent differences in the logic and semantics are merely verbal. In order to take what a theory says about truth and logic seriously, we must take it to have this connection with belief. Thus we can reject the suggestion that accepting a logic and semantics doesn't commit you to anything in the theory of mind. However, I've argued that it needn't always commit you to a particular theory of mind (which might be the view seen in Williams 2014). Whilst a logico-semantic position places some restriction on the theory of mind, some may remain consistent with more than one option. I suggest that supervaluationism and degree theories both fall into this category. This is the first step we need to take to get the experimental approach going. As a logico-semantic theory places restrictions on the theory of mind, we can move from the semantics and logics proposed by a theory of vagueness to what that theory say about beliefs, and from there to how they'd predict a subject would behave in an experiment.

Theory of mind for supervaluationism and degree theories

The next important step is to develop theories of mind for different theories of vagueness; to determine in what way particular theories of vagueness are constrained in what they say about belief. When considering a theory of belief for theories of vagueness, the main

questions arise around the non-classical features of the model. We can associate the familiar notions of (full) truth and (full) falsity with full belief and full disbelief, but non-classical notions like dialetheia, indeterminacy and partial truth are more complicated. A theory of mind for supervaluationism requires considering options for the rational attitude towards indeterminacy. I argue that the semantics and logic rule out both full belief and partial belief, but that at least disbelief and a suspension of belief are both possibilities (I suggest we understand suspension of belief as a kind of meta-cognitive attitude). But out of these two, a suspension of judgement fits most comfortably with the supervaluationist picture; one where it is true of every person that they are tall or not tall, but for some it is indeterminate which due to the ambiguity of the extension of 'tall'. The traditional supervaluationist view of semantic indecision (clearly seen in Fine 1975, for instance) points most clearly most clearly to suspending judgement on indeterminate cases.

Theory of belief for degree theories require understanding the cognitive significance of partial truth. I suggest that the right direction to go with this is with sorts of partial belief. Suggesting other types of attitude like disbelief or suspension of judgement as the rational attitude to partial truths undermines the significance of partial truth. It's not clear what we gain from an infinite range of partial truth values, if we end up with a theory of mind which we can get from a paracomplete or supervaluationist model. Going forward with partial belief, there are at least two different models. On the one hand we can take rational belief to be a precise degree of belief, which matches the degree of truth of a vague proposition (Smith 2010, 2014). On the other hand, we might take it that there is some indeterminacy around degree of belief an agent invests in a vague proposition (MacFarlane 2010). In each case, understanding belief to connect to a strength of tendency to act, in the former case we expect agents to always have the same strength of tendency to act towards a proposition, whereas in the second we can allow that it might waver between a certain range on different occasions. Whilst both give a consistent option, which connects in a meaningful way with action (in contrary to the arguments of Keefe 2016), I argue that MacFarlane's theory is preferable, and that it is more plausible that agents would have an indeterminate kind of belief, opposed to the same very precise degree of belief as each other (and themselves over different occasions).

I've argued that philosophers should take notice of experiments; if they're done well then they can inform the debate on vagueness. Whilst recent studies have shortcomings, this thesis also starts to fill in the picture of how to go about the experimental approach in a better way.

We need a good argument that we can connect theories of vagueness with cognitive attitudes, as well as a theory of what these attitudes are for each theory. I have provided such an argument, as well as suggested ways of filling out the theory of mind for two influential theories of vagueness. I will now reconsider the results of the experiments discussed in chapter 3 in light of my conclusions, as well as consider some other suggestions to further the experimental approach to vagueness.

1i. Supervaluationism in the experimental context

Let's start with supervaluationism. I've argued that a theory of mind for supervaluationism should include a notion of suspension of judgement, to take as the rational attitude towards borderline cases and indeterminant propositions. I suggested we characterise this suspension of judgement as a meta-cognitive attitude:

Meta-cog:* Subject S suspends judgement on P at time t iff S neither believes nor disbelieves P, and believes that she cannot make a judgement on P, at time t.

So I suggest that according to supervaluationism, agents believe the supertrue, disbelieve the superfalse, and neither believe nor disbelieve the indeterminate, and rather believe they cannot make a judgement. Thus in an experimental setting, from this perspective, we shouldn't expect to see participants assenting to either P or ¬P, when P is a borderline case. I think we should expect to see them select an option like 'can't tell', which allows them to express they don't want to make a judgement either way. This matter is slightly complicated by the tautological status of the law of excluded middle in the supervaluationist logic. As 'P v ¬P' is necessarily true, the rational attitude to take to it is belief, which is the attitude we can take the supervaluationist model to predict. Thus even though agents shouldn't believe or disbelieve P and ¬P, they shouldn't dissent from 'P v ¬P', or assent to '¬(P v ¬P)'.

How well does the data from previous experiments align with this interpretation of supervaluationism? Of the studies I considered, Alxatib and Pelletier's (2012) gives the most straightforward data to analyse from this perspective, as their study involved assessing the alethic status of propositions about a borderline case. There participants were asked to judge claims about a man labelled '#2' who was intended to be a borderline case of tallness, with 'true', 'false' or 'can't tell'. The propositions were the following:

#2 is tall

#2 is not tall

#2 is both tall and not tall

#2 is neither tall nor not tall

In chapter 3 I was critical of Alxatib and Pelletier's suggestion of what a supervaluationist should expect to see in the study. They claim that reporting '#2 is neither tall nor not tall' (which I refer to as *neither*) indicates a 'gappy' response from the participants. I argued at the time that it's not so clear why this should be the case, given that the law of excluded middle is tautological in the theory. I've now argued that given that belief aims at truth, and LEM is always necessarily true, supervaluationists should predict participants to assent to it, as it is the rational attitude to take according to their theory (at least without some further arguments about speakers wouldn't recognise its truth). Furthermore, they should expect participants not to assent to *neither*, given that it's necessarily false.

Based on the theory of mind I've discussed for supervaluationism, the set of responses most easily explained by the theory would be for participants to report that they 'can't tell' whether '#2 is tall' and '#2 is not tall'. This is because both are indeterminate (given that #2 is a borderline case), and thus the appropriate attitude is a suspension of judgement. The supervaluationist model predicts that participants would report that *both* and *neither* are false, as both are necessarily false.

The actual patterns seen in the results do not mirror this prediction. The vast majority of participants claim that '#2 is tall' is true (44.7%) or false (40.8%), and claim '#2 is not tall' is either true (25%) or false (65%) (Alxatib and Pelletier 2012: 308-9). Only 9.2% of participants used the 'can't tell' option in response to '#2 is tall' and only 7.9% for '#2 is not tall'. Based on the theory of mind developed here for supervaluationism, we should expect to see many less participants describing an indeterminate proposition as either true or false, and should expect many more to use an option like 'can't tell', which can reflect a suspension of judgement on the matter. For the complex propositions, 44.7% reported that *both* was true, and 40.8% reported it was false, 53.9% reported *neither* was true, and 42% reported it was false. Here there is a relatively even split between true/false responses in each case, with no reported statistical significant difference between them. Supervaluationism predicts that *false* responses would be the likely response, given that these complex propositions are necessarily false. The results from the Alxatib and Pelletier study don't support a supervaluationist view of vagueness. The authors do also reach this conclusion, but for the wrong reasons. My

analysis comes from a stronger position, having developed and argued for a connection between theories of vagueness and theories of mind, and having explored theories of mind for supervaluationism, but reaches the same result.

However, this is far from clear evidence against supervaluationism. For one thing, this brings to light some methodological issues that should be addressed. First, some might worry that participants confuse a proposition like '#2 is neither tall nor not tall' with one like '#2 is neither *definitely* tall nor *definitely* not tall'. Whilst the former is false in supervaluationism, the second can be true. There is suggestion that such confusions are possible, for instance Keefe (2000: 164) suggests that speakers might confuse an instance of the law of excluded middle, such as '#2 is tall or not tall', which cannot be denied on the theory, with a claim like '#2 is definitely tall or definitely not tall', which can be denied. If this is right, then a supervaluationist could point to participants reporting *neither* is true and claim that it is in line with supervaluationism.

Whether speakers do confuse such claims is an empirical question, and the only study (I'm aware of) that investigates this kind of confusion suggests there is no significant phenomena here. Serchuk et al (2011) gave participants a verbal description of a borderline case for the vague predicates 'rich' and 'heavy', such as 'Susan is somewhere between women who are clearly rich and clearly non-rich', and then compared levels of assent to 'Susan is rich (heavy)' and 'Susan is definitely rich (heavy). They find no statistical difference in the way participants respond to the two claims (following appropriate statistical methods I should add!), and take this to be evidence against the suggestion that speakers might confuse those two claims. But there are plenty of reasons to be dissatisfied with that conclusion, most of all that the participants were given a list of twelve sentences to judge which features both 'Susan is rich' and 'Susan is definitely rich', likely highlighting a difference between the two.

Serchuk et al claim they can evade this worry with a second study, using a Bonini et al type design, where separate groups were asked to judge the upper and lower boundaries for 'x is F' and 'x is definitely F' for the vague predicates 'old', 'long' and 'tall'. The suggestion is that if there is confusion between these claims, then there shouldn't be a significant difference between the average height claimed to be the largest someone can be without being considered 'tall' and the largest someone can be without being considered 'definitely tall'. The study compares the 'largest judgers' (i.e. the lowest value where the predicate still truly applies) and 'smallest judgers' (i.e. the greatest value where the predicate still falsely

applies) for each adjective, giving six comparisons overall. In only two of these six cases was there a statistically significant difference between the average value given for ‘x is F’ and ‘x is definitely F’ (the average values given by the ‘largest judgers’ for ‘old’ and ‘long’). In the other four cases there was no significant difference in the average values. Serchuk et al conclude from this that at least the suggestion that speakers confuse the relevant propositions is not fully generalisable, but this is a weak suggestion. That in the majority of cases here speakers do seem to treat ‘x is F’ and ‘x is definitely F’ similarly gives at least some prima facie support for the suggestion that there is possible confusion.

This study alone is far from conclusive, and clearly further research is required. A first step would be to recreate a study like the first of Serchuk et al’s but with *independent groups*. That is, compare the way speakers respond to ‘x is F’ and ‘x is definitely F’ for borderline cases, but with independent groups of participants judging each proposition; this would avoid drawing attention to a difference between the two by asking participants about both in the same study. Further attention also needs to be paid to complex propositions, opposed to the simple propositions used by Serchuk et al. Here we’re primarily interested in whether a supervaluationist could claim speakers assenting to the negation of the law of excluded middle for borderline cases is in support of their view, due to a confusion between ‘¬(P v ¬P)’ and ‘¬(DP v ¬DP)’. To progress further with this issue, we should be comparing the way speakers respond to these claims respectively. With this kind of research, we could make progress and support a case for confusion in these studies. But until that point, we should take participants dissenting from law of excluded middle to run counter to a supervaluationist theory of mind.

A second issue to be considered is the role of ‘can’t tell’ type options in studies like Alxatib and Pelletier’s. As noted, the ‘can’t tell’ option was very rarely used in their study, which is at least initially surprising, as our standard understanding of borderline cases might make it seem this would be a desirable option for participants to use to describe them. Alxatib and Pelletier note this also, wondering if participants didn’t use the option as it represented something like a ‘regrettable lacuna’ in their knowledge. A related point is that the majority of participants were classical in their judgements, with 63% taking one of ‘#2 is tall’ and ‘#2 is not tall’ to be true, and the other false. It may be that participants have a bias towards trying to remain classically consistent in their judgements (as is also considered elsewhere, see Egré and Zehr 2016), and prefer to avoid options like ‘can’t tell’. If this is a genuine phenomenon, then this is problematic, given the characterisation of a supervaluationist theory

of mind based on a suspension of judgement. As supervaluationism predicts participants will take a meta-cognitive attitude towards borderline cases, a belief that they cannot make a judgement either way, for this to be seen in experimental conditions, it's important there aren't confounding variables or biases that stop participants from answering in the relevant patterns. More research can be conducted on the resistance seen to 'can't tell' here, and ways to reduce any possible bias against it. This could involve directly pointing out the possibility of abstaining as a reasonable response to a borderline case, to relieve any pressure participants feel to avoid such 'regrettable lacunas' in their knowledge. Attention should also be paid to other phrases that could be used to capture a suspension of judgement, 'I cannot make a judgement' or 'it isn't possible to tell', for example.

Avenues of research such as these may be important to investigating supervaluationism from an experimental perspective. It may well be that participants don't confuse complex propositions with similar ones containing a 'definitely' operator, and don't stay away from options like 'can't tell' due to unrelated biases. Rather it may be that these sorts of patterns seen in Alxatib and Pelletier are just part of the competent use of vague language, and that this usage *isn't* well modelled by supervaluationism. However, if we can start to identify suspensions of judgement in the face of borderline cases, as part of a competent response to vagueness, then supervaluationism will begin to look like an attractive model. But in any case, extensive further research is required to start making progress along these lines.

1ii. Degree theories in the experimental context

In chapter 6, I argued for a theory of mind for the degree theorist that involved understanding attitudes towards borderline cases as a sort of indeterminate partial belief, modelled by a probability distribution. The distribution represents the likelihood that an agent takes each possible degree of partial belief in a proposition, understood as a strength of tendency to act. How does this play out when it comes to behaviour in an experimental setting?

There are two ways that a partial strength of tendency to act might be seen. First, we could expect agents to give responses that indicate partial confidence. For example, if a participant has a 0.5 partial belief in proposition P, we might see them assent to something like 'partly agree that P, partly disagree that P', or something like 'P is half true'. Second, we could expect a 0.5 belief in P to be seen by agents acting as if P 50% of the time. Over a population, we would then expect there to be a roughly 50/50 split over those that act as if P, and act as if

not P. I suggest that degree theoretic semantics should expect to see participants indicating partial confidence in partially true vague propositions, or a kind of disagreement visible over the whole population.

Importantly, we can expect variation in a single agent's responses to the same vague stimulus across different occasions, relative to both of the ways a partial belief could be seen in a study. In the first case, as on my suggested theory of mind agent's don't always have a particular precise partial belief to a borderline case, but rather their attitude can vary between a range, with some more likely than others. This means on different occasions, particular strength of tendency to act can vary, resulting in differences in the response pattern given. In the second case, when strength of tendency to act is understood as a frequency of action, participants can be expected to vary, as a 0.5 partial belief in P means acting as if P only half of the time.

It is interesting to point this out, as there is some evidence of participants varying in their responses to borderline cases. For instance, it was famously seen in McCloskey and Glucksberg (1978), who are interested in whether some categories such as 'sport' and 'science' are fuzzy-bounded. They have a range of items, some which are clear positive, and negative cases, and some which are meant to be a kind of borderline case (for instance, football, watching TV and chess are respective examples for the category 'sport'). They gave participants a list of items, asking if they were in some relevant category, then gave them the same test a month later. Interestingly for items that were meant to be borderline cases, relatively high inconsistency was found between the first and second test. For 'level 4' borderline cases, participants changed their response 22% of the time, and 18% at 'level 5' (level 1 items clearly not in the relevant category and level 9 items clearly are).

Inconsistency is also found in Hampton et al (2012). In this piece a number of experiments are carried out in which participants were asked to judge sentences of different natures. Most relevant here is category statements, such as 'olive is a fruit', a number of which were meant to be borderline style cases. In one condition, participants were given two options; 'true' and 'false'. In the second condition, participants were given the options '100% sure it's true, 'not 100% sure either way' and '100% sure it's false'. Again, participants completed the questionnaire on two different occasions. In one experiment Hampton et al found that participants were consistent in their responses only 82% of the time in the two-option condition. When they had three options, participants were less consistent; 77% of the time if

they'd answered '100% sure it's true', 73% of the time if they'd answered '100% sure it's false', and 54% of the time if they're answered 'not 100% sure either way'. It was found there was no significant difference in consistency between the two conditions. In a second experiment which focussed on borderline category statements, with over two thirds being identified as borderline cases, there was on average 83.2% consistency in the two-response condition, and on average 83.6% consistency in the three-response condition.

Here we do see some evidence of inconsistency in the way speakers judge borderline cases. This is hard to explain from a supervaluationist view point; speakers are expected to suspend judgement on each occasion, so shouldn't vary. This is more easily explained from the perspective of degree semantics, understanding participants to have partial beliefs. In these studies we see roughly the same percentage of speakers judge a borderline case one way or another, but each time it is a different set of speakers giving those responses, as can be expected given the variances that can come with partial belief.

These particular studies also show the between-subjects disagreement that can be expected given one understanding of strength of tendency to act. For instance, in McCloskey and Glucksberg, disagreement about the classification of an object was statistically significantly higher for the borderline cases than the more clear-cut cases, peaking for objects that had level 4 typicality, where there was on average a 64% to 36% split (1978: 465). Exactly what percentage split should be expected based on the degree theoretic theory of mind depends on the degrees of the relevant partial belief. Here perhaps we'd expect to see a more even split, but what's most important is that disagreement between subjects is significantly higher for borderline cases, and this is a result that is consistent with the view of partial beliefs based on degree semantics.

No studies considered earlier in the thesis investigate within-subject consistency or give options to express partial confidence. However, we can see patterns in Alxatib and Pelletier's results that are arguably amenable to the degree theoretic approach, including appropriate levels of between-subject disagreement. In chapter 3, we saw that Alxatib and Pelletier reject a degree theoretic interpretation of their data. They suggest that even though the medium levels of assent to contradictions is consistent with the 0.5 degree of truth they'd receive according to degree semantics, a degree theorist couldn't explain the overall patterns in the data, in particular the responses to the simple propositions given the responses to the complex ones.

I was critical of this in chapter 3, primarily because of the lack of a well-developed theory of mind for degree semantics. In this thesis I have developed a better theory of the connection between logico-semantic theories, and theory of mind, and have considered some cognitive theories for degree theoretic accounts of vagueness in depth. We can now reassess the data from the perspective of the theory of mind I've considered, and we can see that the data could be explainable by degree theories.

First, recall the result Alxatib and Pelletier have in mind when they reject degree theoretic accounts is one where participants assent to '#2 is tall and not tall' *(both)*, but reject '#2 is tall' and reject '#2 is not tall'. But as I've argued, that is not a statistically likely result, and was seen from only 11 of the 79 participants, so needn't be paid much attention. What should we expect to see from the results in Alxatib and Pelletier based on the theory of mind I've considered? As Alxatib and Pelletier don't give any options to express a partial confidence (unless we count 'can't tell', but as discussed there are problems with that option), we should expect to see disagreement between speakers over the borderline cases, and disagreement there being significantly higher than for the clearer cases.

This is the kind of pattern we see emerge. There was disagreement over the borderline proposition '#2 is tall', with 44.7% claiming it was true, 40.8% claiming it was false. Again with the borderline '#2 is not tall', 25% claim it is true, 65% claim it is false (Alxatib and Pelletier 2012: 308-9). This disagreement is much higher than it is for the other examples in the study, with almost all participants agreeing that the 6'2" and 6'6" men were tall, and the 5'4" and 5'7" men were not tall (2012: 307-8). The disagreement seen for the borderline cases is consistent with the suggestion that speakers had a middling partial belief, unlike in the other clear cases. Similar things can be said about the response to the complex propositions *both* and *neither*. As noted, 44.7% of participants reported that '#2 is tall and not tall' was true, and 40.8% reported it was false, 53.9% reported '#2 is neither tall nor not tall' was true, and 42% reported it was false. These classical contradictions can reach up to 0.5 true on degree semantics, which models the levels of assent here quite well; participants are in disagreement over the classification of the propositions, and quite evenly split. The levels of assent and disagreement to these propositions is also highest for the borderline case, and is relatively low for the others (2012: 314), which fits well with the degree theoretic model, given that the truth value of those proposition peaks for the borderline case.

It should also be noted that here the degree theoretic account models the assent to contradictions much better than a dialetheist theory (e.g. Priest's LP, as defended as an account of vagueness in Webber et al 2014), or Alxatib and Pelletier's own pluralistic account. On these theories contradictions are fully true, and thus assent should be expected to be much higher than the 44.7% seen in Alxatib and Pelletier's study. Alxatib and Pelletier's own novel account, '#2 is tall and not tall' is true of individuals that are sub-tall and either borderline or super-not-tall, which can only consistently be the borderline cases. But as #2 is a borderline case, '#2 is tall or not tall' is true. This model should expect to see much higher levels of assent then they record, and much less disagreement. The degree theoretic theory of mind explains this much more straightforwardly; as *both* is borderline true, we should expect to see middling assent and relatively even disagreement.

A similar point can be made with other studies. For instance, Ripley (2011) presents evidence of some assent to contradictions, and is most interested in a dialetheist or contextualist interpretation. In the study, participants were shown 7 different pairs consisting of a circle and a square, pair A-G. In each pair the square is on the far right of the image, but the placing of the circle differs. In pair A, the circle is on the far left, far away from the circle. In the following pairs the circle is shown to be progressively closer, until it is right next to the square in pair G. The idea is that pair A is a clearly false instance of the sentence 'the circle is near the square' whilst pair G is a clearly true instance of that same sentence, and some of the pairs in between are borderline cases. Participants were each asked to indicate their level of agreement to a classical contradiction a scale of 1 (full disagreement) to 7 (full agreement) to each of the pairs. Ripley found that the highest mean level of agreement to one the pairs was 4.1 to pair C, whilst the mean peak level of agreement was 5.4. But a model according to which contradictions are fully true cannot easily explain why the levels of agreement are so low here. A degree theoretic account can explain this: there is middling agreement because the contradictions are only partial truths.

So we can see, that armed with our well-developed theory of mind for the degree theorist, we can make claim that Alxatib and Pelletier's results do support a degree account (along with some others, McCloskey and Glucksberg 1978, Ripley 2011). Clearly, this account has a better claim to explain the data than supervaluationism does; the levels of assent look better modelled by partial beliefs than by suspension of judgement.

There is need for more research still. Here I have focussed on understanding strength of tendency to act as representative of a frequency of action, but this is because there is limited research into alternative response types, that might express a partial confidence. Ripley (2011) does give us an example of this, with a scale on which to report a degree of acceptance, and there we saw participants were likely to give a middling response to a borderline case. More research should be conducted on these sorts of scales. Further attention should also be given to 'partly agree, partly disagree' type responses, as was the case with 'can't tell' type answers. It is reasonable that a degree theorist might expect these kinds of answers to be given in an experimental setting, but there may be similar issues relating to a bias against seeming ignorant, as we arguably saw from participants in Alxatib and Pelletier. These avenues can help extend the investigation into a degree theoretic model of vagueness.

I don't think that the data from Alxatib and Pelletier alone is sufficient for drawing any sort of conclusions, but we see that on back of the theoretical work in this thesis, we can at least make a claim to interpret it from a better position than the authors. With a well-developed notion of theory of mind for supervaluationism and degree theories, we can see that the degree theorist has a better claim to explaining the response types seen than the supervaluationist: they appear more consistent with a notion of partial beliefs than of suspensions of judgement. This approach sets a good example to future experimentalists. Understanding our theories of vagueness as descriptive theories, the connections between the logic and semantics and theory of mind, we can go onto to develop theories of mind for theories of vagueness which can be compared with experimental results.

Bibliography

Alexander, J., Mallon, R., and Weinberg, J. M. 2010. ‘Accentuate the negative’. *Review of Philosophy and Psychology* 1: 297–314

Alxatib, S. and Pelletier, F. J. 2011. ‘On the psychology of truth-gaps. In *Vagueness in Communication*, ed. R. Nouwen, R. van Rooij, U. Sauerland, and H. C. Schmitz. VIC 2009. Lecture Notes in Computer Science, vol 6517: Springer

Alxatib, S. and Pelletier, F. J. 2012. ‘The psychology of vagueness, borderline cases and contradictions’. *Mind and Language* 26(3): 287–326

Bergmann, M. 2005. ‘Defeaters and higher-level requirements’. *The Philosophical Quarterly* 55: 419–436

Black, M. 1937. ‘Vagueness: An exercise in logical analysis’. *Philosophy of Science* 4(4): 427-455

Bonini, N., Osherson, D., Viale, R. and Williamson, T. 1999. ‘On the psychology of vague predicates’. *Mind and Language,* 14(4): 377-393

Borel, É. 1907. ‘Sur un paradoxe économique: le sophisme du tas de blé et les vérités ´ statistiques’. *Revue du Mois* 4: 688-699

Braisby, N., Franks, B. and Hampton, J. 1996. ‘Essentialism, word use, and concepts’. *Cognition* 59: 247–274

Brown, S., Attardo, S. and Vigliotti, C. 2014. *Understanding Language Structure, Interaction, and Variation: An Introduction to Applied Linguistics and Sociolinguistics for Nonspecialists*. Third ed: Michigan ELT

Buckwalter, W. 2010. ‘Knowledge isn’t closed on Saturdays: a study in ordinary language’. *Review of Philosophy and Psychology* 1: 395–406

Burgess, A. and Plunkett, D. 2013. ‘Conceptual ethics I’. *Philosophy Compass* 8(12): 1091–1101

Cappelen, H. ‘X-Phi without intuitions? In *Intuitions*, ed. A. Robert Booth and D. P. Rowbottom, 2014: Oxford University Press

Chisholm, R. 1976. *Person and object*: Open Court

Cobreros, P. 2010. ‘Paraconsistent vagueness: a positive argument’. *Synthese* 183: 211-227

Cobreros, P., Egré, P., Ripley, D., van Rooij, R. 2012. ‘Tolerant, classical, strict’. *Journal of Philosophical Logic* 41: 347-385

Crawford, S. 2004. ‘A solution for Russellians to a puzzle about belief’. *Analysis* 64: 223–229

Culbertson, J. and Gross, S. 2009. ‘Are linguists better subjects?’. *British Journal of Philosophy of Science* 60(4): 721–736

Dummett, M. 1975. ‘Wang’s paradox’. *Synthese* 30: 301-24

Edgington, D. 1997. ‘Vagueness by degrees’. In *Vagueness: A Reader,* ed. R. Keefe and P. Smith, 1997

Egré, P. 2017. ‘Vague judgement: a probabilistic account’. *Synthese* 194: 3837–3865

Egré**,** P., de Gardelle, V. and Ripley, D. 2013. ‘Vagueness and order effects in color categorization’, *Journal of Logic, Language and Information*, 22(4): 391-420

Egré, P. and Klinedinst, N. 2011. *Vagueness and Language Use*: Palgrave Macmillan

Egré, P. and Zehr, J. 2016. ‘Are Gaps preferred to Gluts? A closer look at borderline contradictions’. Forthcoming in *The Semantics of Gradability, Vagueness, and Scale Structure - Experimental Perspectives,* ed. E. Castroviejo, G. Weidman Sassoon, and L. McNally: Springer

Field, H. 2009. ‘What is the normative role of logic’. *Proceedings of the Aristotelian Society* 83: 251–68

Field, H. 2009b. ‘Pluralism in logic’. *The Review of Symbolic Logic* 2(2): 342–359

Field, H. 2010. ‘This magic moment: Horwich on the boundary of vague terms’. In *Cuts and Clouds: Vagueness, its Nature and its Logic*, ed. R. Dietz and S. Moruzzi, 2010: Oxford University Press

Fine, K. 1975. ‘Vagueness, truth and logic’. *Synthese* 30: 265-300. Reprinted in Keefe and Smith 1997

Frege, G. 1879. *Begriffsschrift. Eine der rithmetischen nachgebildete Formelsprache des reinen Denkens*: Verlag von Louis Nebert

Frege, G. 1903. *Grundgesezte der Arithmetik,* vol. II: Hermann Pohle

Friedman, J. 2013. ‘Suspended judgement’. *Philosophical Studies* 162: 165–181

Friedman, J. 2015. ‘Why suspend judging?’. *Noûs* 51(2): 302-326

Gettier, E. K. 1963. ‘Is justified true belief knowledge?’ *Analysis* 23: 121-123

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., and Cohen, J. D. 2001. An fMRI investigation of emotional engagement in moral judgment. *Science* 293: 2105-2108

Hampton, J. A. 2007. ‘Typicality, graded membership, and vagueness’. *Cognitive Science* 31(3): 355-384

Hampton, J. A., Aina, B., Andersson, J.M., Mirza, H. and Parmar, S. 2012. The Rumsfeld Effect: the unknown unknown. *Journal of Experimental Psychology: Learning Memory & Cognition* 38: 340–355.

Hampton, J. A. and Passanisi, A. 2016. ‘When intensions do not map onto extensions: Individual differences in conceptualization’. *Journal of Experimental Psychology: Learning Memory and Cognition* 42(4): 505–523

Hansen, N. 2015. ‘Experimental philosophy of language’. *Oxford Handbooks Online*

Hansen, N. and Chemla, E. 2015. ‘Linguistic experiments and ordinary language philosophy’. *Ratio*. 28: 422-445

Harman, G. 1986. *Change in View*: M.I.T. Press

Horvath, J. 2010. ‘How not to respond to experimental philosophy’. *Philosophical Psychology* 23(4): 447-480.

Horvath, J. and Grundmann, T. (eds.) 2012. *Experimental Philosophy and its Critics*: Routledge.

Hyde, D. 1997. ‘From heaps and gaps to heaps of gluts’. *Mind* 106: 641-660

Hyde, D. 2008. *Vagueness, Logic and Ontology*: Ashgate

Jylkka, J., Railo, H., and Haukioja, J. 2009. ‘Psychological essentialism and semantic externalism: evidence for externalism in lay speakers’ language use’. *Philosophical Psychology* 22: 37–60

Kalmus, H. 1979. ‘Dependence of colour naming andmonochromator setting on the direction of preceding changes in wavelength’. *British Journal of Physiological Optics,* 32(2): 1–9

Kauppinen, A. ‘The rise and fall of experimental philosophy’. *Philosophical Explorations* 10(2): 95 – 118

Keefe, R. 2000. *Theories of Vagueness:* Cambridge University Press

Keefe, R. 2017. ‘Degrees of belief, expected and actual’. *Synthese* 194(10): 3789–3800

Keefe, R. and Smith, P. 1997. *Vagueness: A Reader*: MIT Press

Kim, J. and Sells, P. 2008. *English Syntax: An Introduction*: CSLI Publications

Knobe, J. 2003. ‘Intentional action in folk psychology: an experimental investigation’. *Philosophical Psychology* 16: 309–324

Knobe, J. and Nichols, S. 2008. ‘An experimental philosophy manifesto’. In *Experimental Philosophy,* ed. J. Knobe and S. Nichols 2008: Oxford University Press

Knobe, J. and Nichols, S. (eds.) 2008. *Experimental Philosophy*: Oxford University Press

Knobe, J. and Yalcin, S. 2014. ‘Epistemic modals and context: experimental data’. *Semantics and Pragmatics*. 7(10): 1–21

Krifka, R. 2011. ‘Varieties of semantic evidence’. In *Semantics: An International Handbook of Natural Language and Meaning,* ed. C. Maienborn, K. von Heusinger, and P. Portner 2011: de Gruyter

Kripke, S. 1980. *Naming and Necessity*: Basil Blackwell

Lewis, D. 1986. *On the Plurality of Worlds*: Blackwell

Ludwig, K. 2007. ‘The epistemology of thought experiments’. *Midwest Studies in Philosophy XXXI*: 128–159

MacFarlane, J. 2004 (unpublished). ‘In what sense (if any) is logic normative for thought?’. Retrieved from <http://johnmacfarlane.net>

MacFarlane, J. 2010. ‘Fuzzy Epistemicism’. In *Cuts and Clouds: Vaguenesss, its Nature and its Logic*, ed. R. Dietz and S. Moruzzi, 2010: Oxford University Press

Machery, E. 2012. ‘Expertise and intuitions about reference’. *Theoria* 73: 37–54

Machina, K. F. 1976. ‘Truth, belief and vagueness’. *Journal of Philosophical Logic* 5: 47-78

Marti, G. 2009. ‘Against semantic multi-culturalism’. *Analysis* 69(1): 42–48

McCloskey, E. and Glucksberg, S. 1978. ‘Natural categories: Well defined or fuzzy sets?’ *Memory & Cognition* 6: 462-472

Mehlberg, H. 1956. ‘Truth and vagueness’. In *Vagueness: A Reader*, ed. R. Keefe and P. Smith, 1997

Mill, J. S. 1884. *A System of Logic*: Longman

Mundici, D. 2006. ‘Bookmaking over infinite-valued events’. *International Journal of Approximate Reasoning* 43: 223–40

Nouwen, R., van Rooij, R., Sauerland, U. and Schmitz, H. 2011. *Vagueness in Communication*: Springer

Nichols, S. and Knobe, J. 2007. ‘Moral responsibility and determinism: the cognitive science of folk intuition’. *Nôus*, 41: 663–685

Paris, J. B. 2001. ‘A note on the Dutch Book method’. Pages 301–306 of: *Proceedings of the Second International Symposium on Imprecise Probabilities and their Applications, ISIPTA*: Shaker

Petrinovich, L. O’Neill, P. 1996. ‘Influence of wording and framing effects on moral intuitions’. *Ethology and Sociobiology* 17: 145-171

Pietroski, P., Lidz, P., Hunter, T. and Halberda, J. 2009. ‘The meaning of ‘most’: semantics, numerosity and psychology’. *Mind & Language* 24: 554–585

Priest, G. 1979. ‘The logic of paradox’. *Journal of Philosophical Logic* 8: 219-241

Putnam, H, 1975. ‘Is semantics possible?’. In *Mind, Language and Reality: Philosophical Papers*, Volume 2, 139–152: Cambridge University Press

Raffman, D. 2014. *Unruly Words*: Oxford University Press

Ramsey, F. P. 1926. ‘Truth and probability’. In *Philosophical Papers*, ed. D. H. Mellor, 1990

Ramsey, F. P. 1990. *Philosophical Papers*: Cambridge University Press. Ed. D. H. Mellor

Ripley, D. 2011. ‘Contradictions at the borders’. In *Vagueness in Communication*, ed. R. Nouwen, R. van Rooij, U. Sauerland and H. Schmitz, 2011: Springer

Ripley, D. 2012. ‘Sorting out the sorites’. In *Paraconsistency: Logic and Applications*, ed. F. Berto, E. Mares, and K. Tanaka 2012: Springer

Ripley, D. 2016. ‘Experimental philosophical logic’. In *A Companion to Experimental Philosophy*, ed. J. Sytsma and W. Buckwalter, 2016: John Wiley and Sons, Ltd

Ritov, I. and Baron, J. 1990. ‘Reluctance to vaccinate: omission bias and ambiguity’. *Journal of Behavioral Decision Making* 3: 263-277

Russell, B. 1923. ‘Vagueness’. In *Vagueness: A Reader*, ed. R. Keefe and P. Smith, 1997

Sainsbury, R. M. 1995. *Paradoxes*: Cambridge University Press

Sauerland, U. 2011. ‘Vagueness in language: The case against fuzzy logic revisited’. In *Understanding Vagueness: Logical, Philosophical, and Linguistic Perspectives*, ed. P. Cintula, C. Fermuller, L. Godo, and P. Hajek, 2011: College Publications

Sauerland, U. and Stateva, P. 2011. ‘Two types of vagueness’. In *Vagueness and Language Use*, ed. P. Egré and N. Klinedinst, 2011: Palgrave Macmillan

Scharp, K. 2013. *Replacing Truth*: Oxford University Press

Schiffer, S. 2000. ‘Vagueness and partial belief’. *Noûs* 34(1): 220 - 257

Schiffer, S. 2003. *The Things We Mean*: Oxford University Press

Serchuk, P., Hargreaves, I. and Zach, R. 2011. ‘Vagueness, logic and use: Four experimental studies on vagueness’. *Mind and Language* 26(5): 540-573

Sextus Empiricus. 2000. ‘Outlines of scepticism’. *Cambridge Texts in the History of Philosophy*: Cambridge University press. Translated by Julia Annas and Jonathan Barnes.

Shapiro, S. 2006. *Vagueness in Context*: Oxford University Press

Smith, N. J. J. 2003. ‘Vagueness by numbers? No worries’. *Mind* 112: 283-290

Smith, N. J. J. 2008. *Vagueness and Degrees of Truth*: Oxford University Press

Smith, N. J. J. 2010. ‘Degree of belief is expected truth value’. In *Cuts and clouds: Vagueness, its nature and its logic*. Ed. R. Dietz and S. Moruzzi, 2010: Oxford University Press

Smith, N. J. J. 2014. ‘Vagueness, uncertainty and degrees of belief: two kinds of indeterminacy—One kind of credence’. *Erkenntnis* 79: 1–18

Sosa, E. 2007. ‘Experimental philosophy and philosophical intuition’. *Philosophical Studies* 132(1): 99-107

Spranca, M., Minsk, E. and Baron, J. 1991. ‘Omission and commission in judgment and choice’. *Journal of Experimental Social Psychology* 27: 76-105

Sprouse, J., Schütze, C. T. and Almeida, D. 2013. “A comparison of informal and formal acceptability judgments using a random sample from *Linguistic Inquiry* 2001–2010”. *Lingua* 134: 219–248

Sturgeon, S. 2010. ‘Confidence and coarse-grained attitudes’. In *Oxford studies in Epistemology (Vol. 3)*: Oxford University Press

Swain, S., Alexander, J., Weinberg, J. M. 2008. ‘The instability of philosophical intuitions: running hot and cold on truetemp’. *Philosophy and Phenomological Research* 76: 138-155

Tye, M. 1994. ‘Sorites paradoxes and the semantics of vagueness’. *Philosophical Perspectives,* 8: *Logic and Language*¸ ed. J. E. Tomberlin: Ridgeview

Van Fraassen, B. C. 1998. ‘The agnostic subtly probabilified’. *Analysis* 58: 212–220

Varzi, A. 2001. ‘Vagueness, logic and ontology’. *The Dialogue* 1: 135-154

Varzi, A. 2007. ‘Supervaluationism and its logics’. *Mind* 116: 633-676

Webber, Z., Hyde, D., Ripley, D., Colyvan, M. and Priest, G. 2014. ‘Tolerating gluts’. *Mind* 123: 813-828

Weinberg, J. M., Nichols, S. and Stitch, S. 2001.’Normativity and epistemic intuitions’. *Philosophical Topics* 29: 429-460

Williams, J. R. G. 2011. ‘Degree supervaluational logic’. *Review of Symbolic Logic* 4(1): 130-149

Williams, J. R. G. 2012. ‘Gradational accuracy and non-classical semantics’. *Review of Symbolic Logic* 5(4): 513-537

Williams, J. R. G. 2014. ‘Non-classical minds and indeterminate survival’. *Philosophical Review* 123(4): 379-428

Williamson, T. 1994. *Vagueness*: Routledge

Wright, C. 1976. ‘Language-mastery and the sorites paradox’. In *Truth and Meaning: Essays in Semantics,* ed. G. Evans and J. McDowell: Clarendon Press

Zadeh, L. A. 1965. ‘Fuzzy sets’. *Information and Control* 8: 338-53