"Bo," said Jack, "I'm not sure we should be eating the Professor's allergy medicine. There might be all kinds of side effects." Bo popped a last pill in her mouth. "I assumed we would be ok," she said, and chewed thoughtfully. ***"Now that you mention it, I wonder…"*** She munched a moment longer, then swallowed. "Well, I suppose now we'll find out." And indeed they did: seconds later Jack noticed Bo's hat lifting off her head, carried by ears growing steadily longer and covered with fine hair. ***Awareness,*** the sting of hindsight, came over them. "***Attention,*** donkeys," blared the Professor's voice. "***Assumption*** makes an ass of you and me," he trumpeted —Bo squinted down her muzzle and decided not to correct him— "and you should be grateful for the small mercy that I am not American." He sneezed suddenly. "Ach, my allergies, I must get you out of here and put you to work." He called in a short hairy man (*moi*) carrying two rope halters and a bag of carrots. The Professor made introductions: "***Tikitu de Jager*** will train you in pragmatics and ploughing." Waggling the carrots enticingly I led them away.

"Now that you mention it, I wonder…"

Awareness, Attention, Assumption

ILLC Dissertation Series DS-2009-10

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

For further information about ILLC publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Science Park 904
1098 XH Amsterdam
phone: +31-20-525 6051
fax: +31-20-525 5206
e-mail: illc@science.uva.nl
homepage: http://www.illc.uva.nl/

"Now that you mention it, I wonder…"
Awareness, Attention, Assumption

Academisch Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam
op gezag van de Rector Magnificus
Prof. dr. D. C. van den Boom
ten overstaan van een door het college voor
promoties ingestelde commissie, in het openbaar
te verdedigen in de Aula der Universiteit
op dinsdag 15 december 2009, te 12.00 uur

door

Samson Tikitu de Jager

geboren te Takaka, Nieuw-Zeeland.

Promotiecommissie

Promotor: Prof. dr. ir. R. J. H. Scha
Co-promotor: Dr. ing. R. A. M. van Rooij

Overige commissieleden:
Prof. dr. H. Kamp
Prof. dr. F. Landman
Prof. dr. M. J. B. Stokhof
Prof. dr. F. J. J. M. Veltman
Dr. O. Board

Faculteit der Geesteswetenschappen

Printed and bound by Ipskamp Drukkers.

ISBN: 978–90–5776–201–7

Contents

Acknowledgements

One doesn't write a book every day, and I probably won't write another anytime soon, so I take the opportunity to publically discharge some longstanding debts of gratitude. First under this heading must come my parents, Hennie, Paul, and Brian: for years of love and support, and for (in their various ways) setting high standards. Ethically, aesthetically, spiritually, and politically, they are still asking me tough questions (whether they realise it or not), and I am still learning from trying to answer them.

In my academic career I seem to have been handed along a chain of mentors, the lack of any of whom would have prevented this dissertation from appearing. Alistair Knott, teaching at Otago University in New Zealand, got me excited about the connections between artificial intelligence and language, and paid me good money for bad programming: my first research position. Without Ali I probably would have settled for a programming career (terrifying thought). Hans van Ditmarsch, also at Otago, showed me the muddy children problem and pointed me at the Master of Logic programme in Amsterdam. This in turn introduced me to the ILLC and the rich interdisciplinary approach to logic favoured here; I suppose there are few other places in the world where a Java programmer could pick up the basics of analytic philosophy of language while writing a masters thesis on algorithm analysis and set theory.

My MSc thesis supervisor Benedikt Löwe is also an essential link in the chain: he is almost singlehandedly responsible for the fact that I neither starved to death in Amsterdam nor had to return penniless to New Zealand after missing the application deadlines for every possible scholarship at the end of my first year. Benedikt found me part-time typesetting work that kept the wolf from the door, and I presume it was he who put my name forward as a possible recipient of the ILLC Scholarship; without that financial support I could not have stayed in the Netherlands to finish my MSc. He also encouraged me in my typographic interests, which (I'm afraid) will remain with me long after the set theory has faded.

Henk Zeevat was my official academic mentor for about three weeks at the beginning of the MoL programme, and thereafter a friend who incidentally was supposed to give me advice when I needed it. Most of the time I didn't need it (as a substitute he taught me to ski, and —less successfully— to ice-skate), but he came through in the best possible way near the end of those first two years.

When I had no attention for anything but writing up, he pointed out to me an upcoming PhD position at the ILLC; when I dragged my heels he pushed me into making an application; and he gave me what must have been a glowing reference (since I had never taken classes with the fellow offering the position, and indeed not many in the area of research he was pursuing).

Which brings me to Robert van Rooij, supervisor of this dissertation, whose PhD student position I was applying for. I owe Robert thanks for many things, but the first (chronologically speaking) is that he told me at the start of my interview, "We've already had a candidate with a very strong application, so I have to tell you that you probably won't get the job." This left me with nothing to prove, so to speak, and so we ended up having a wing-ding argument about (if I remember rightly) whether game-theoretic models are suitable for representing real human cognition; since I didn't have to impress him to land the job I anyway wouldn't get, I loosened up, got opinionated, and disagreed with most of what he had to say (probably not on particularly strong grounds, in retrospect, but in any case *with feeling*). Apparently this made an impression: when the strong candidate accepted another offer, I was second in line and took the position.

Robert's project ran for four years and funded two PhD positions; the other position was taken by Michael Franke, another MoL graduate. Micha has gone from competitor (when all we knew was that there were two positions and who-knew-how-many applicants) to officemate to friend and collaborator. He is also my yoga teacher. We have bounced ideas off each other so hard they hurt; we have stood on our heads together on the office floor; we have written joint papers and critiqued each others' work; we have hollered and whooped down a Polish ski-slope.

I am not an easy person to work with. I am stubborn, quick to jump to conclusions, and rhetorically flexible enough to defend them in the face of near-overwhelming evidence. My voice gets louder as I retreat to shakier ground. Micha is the very opposite: quiet-voiced, unfailingly careful in both his rhetoric and his conclusions, and always considering the feelings of whoever he is debating with. He is also, though, in his own way stubborn: on the many occasions when his quietly-held opinions have been right and my loudly-asserted ones have been wrong, he has patiently reeducated me. I am enormously grateful that he has stuck to doing this, even outside our explicit collaborations. Parts of this dissertation grew out of joint work with Micha, but also much that does not explicitly bear that notice is the result of conversations with him or (as in the chapter on Sobel sequences) the working out of ideas that originate with him.

Micha is also largely responsible for my surviving the last months of writing up without lasting psychological damage: without the weekly yoga classes my

state of mental health would have been *much* worse. He even gave me a mat, to encourage me to practise daily at home.

Those last months wouldn't have been so stressful if I hadn't changed the direction of my research, and the topic of the dissertation, rather further through the project than was really advisable. Robert gets another round of thanks here, for allowing the change of direction (away from his own research interests) and for his support and help especially in the final stages. Robert's style is more like mine than Micha's: loud, quick, and argumentative. Unlike me, he has learned restraint; I am sure he has been as frustrated (on occasion) with my fast-and-loose style as I have been (on occasion) with his, but he has never called me on it. (I, on the other hand, have done — and have had to apologise, when careful reflection has proved me wrong.) He has put up with me telling him that his ideas are stupid, and also with my coming back triumphantly, weeks later, with those same ideas dressed up in alternative notation as my latest contribution to the project.

In the last months and weeks, both he and Remko Scha (my promotor, the professor officially responsible for my PhD confirmation) read far too many drafts and redrafts, giving enormously valuable advice (ranging from "these words are misspelled" through "this definition doesn't make any sense" to "you should rewrite this chapter the other way up", and taking in "why don't you link x with y and call it z, instead of linking x with p and calling it sort-of-but-not-quite c?", as well as the essential "so-and-so already did this/proved this is wrong, read these papers" accompanied by a flurry of references). In other words, I have been blessed with hands-on supervisors, and the thesis is infinitely the better for it. (That it is not still better is of course my responsibility; especially it has suffered from the time pressure caused by my last-minute course change, and —perhaps more importantly— my habit of procrastination. I should also acknowledge that Robert has spent roughly three years warning me about this very danger.)

Besides the academic community, I owe grateful thanks to some Amsterdam folk who have helped me keep a life outside the office.

Ralph, Pippa, Ella and Bea were my prosthetic family on my first arrival in Foreign Parts. Pippa is the twin sister of a friend of one of my mother's neighbours, by reason of which intimate connection the family virtually adopted me: they fed me wholesome meals and invited me to jam sessions and took me out boating and gave me a place to stay when I needed it.

Similarly generous in virtually adopting me were the Dialogue Drinkers: Marijn, Tom, and Frans. These guys are responsible for the first Dutch I learned: they took me to the pub, then refused to speak English or to respond if I used any. ("Dialogue Drinkers" because our first get-together was to celebrate finishing Henk's course Dialogue Systems. Five years later we're still celebrating

the same thing, another lasting effect Henk has had, albeit indirectly, on my Amsterdam experience.)

Stefan Bold shared his house, his library, and his love of sci-fi series television; also his cooking knives and his soup-and-curry expertise.

The Hermeneutic Heideggers get somewhat ambivalent thanks for helping in my various flirtations with depression, alcoholism, and tattooed strangers. I am grateful (wholeheartedly, for the most part) that none of these flirtations bore any fruit, but I'm not sure what balance of blame and credit to assign to the HHers.

A number of people have given me specific suggestions about this project; not all of these have made it into the text, but all have had an influence on how I think about the material. In roughly chronological order (and taking for granted the pervasive influence of Micha, Robert, and in the late phases Remko), I offer grateful thanks for the following help: Edgar Andrade-Lotero raised a tricky problem; Yanjing Wang solved Edgar's problem, and pointed me at valuable related work; Floris Roelofsen likewise brought related work to my attention, and let me argue with him about data semantics — in addition, Floris shared an office with Micha and myself and accepted with our strange behaviours and noise, and my clutter; Jeroen Groenendijk also passed on interesting related work; Emmanual Chemla had a lovely suggestion about "must" which I still have not managed to pin down; Martin Stokhof discussed Wittgenstein with me (I regret that I did not have time to pursue this connection, which I think may go much deeper than the treatment I give it in passing); Maria Aloni and Paul Égré gave me the chance to apply my ideas to their problem, and then put up with my cranky co-authorship; Anton Benz generously invited me to present these ideas at the ZAS in Berlin, where we had an inspiring discussion; Katrin Schulz discussed the cutting-edge version of her theory of counterfactuals with me; and Frank Veltman and Oliver Board both offered comments on draft versions of chapters of the dissertation. Doubtless there are people I have forgotten to mention (my notes for the months immediately preceding my deadline are particularly sparse, while the help I received was anything but), to whom I hereby offer my apologies and thanks.

Finally I would like to offer my love to Olga, and my thanks especially for her support and presence in the penultimate stage of writing (when she took care of me rather as one takes care of a sufferer of senile dementia: reminders to bathe and to eat, preparing meals, making sure I went out wearing a raincoat and not a dressing gown) and for her support and absence in the ultimate stage (which she spent in Greece, leaving me with a guaranteed supply of Greek coffee and the freedom to not clean up after myself for the crucial month).

To all those named and mentioned above: Thank you. Without you, my dissertation would not be what it is; without many of you, it would not be at all.

Chapter 1

Introduction

> The solution was obvious, as obvious as it had seemed insoluble for as long as he hadn't solved it[.]
>
> Georges Perec, *Life A User's Manual*

1 · *An idea too simple to disagree with*

We are surrounded all our lives by myriads of possibilities, strictly speaking perhaps *limitless* possibilities. We, however, are most definitely limited beings; limited certainly in our capacity to understand what possibilities really obtain, but even limited in our awareness of what possibilities *might* obtain. There may be countless unrealised possibilities, but at any given moment we are only aware of a very small subset of them.

If you accept this truism, you are halfway to accepting the argument of this dissertation. The other half consists in the recognition that the semantic structures underpinning our use and understanding of language must reflect this limited awareness; this second half will require a little more justification.

In this chapter I will introduce the core ideas of unawareness and assumption, giving some motivation and describing some existing work that I will build on in the rest of the dissertation. I start with the possible worlds semantics for natural language as defended by Robert Stalnaker; taking his ideas to their logical conclusion leads quickly to an intuitive understanding of the relevance of awareness for formal semantics. In particular, I will raise two kinds of problem for Stalnaker's account: the finegrained individuation of worlds, and cases of worlds that are apparently neither ruled out nor accepted as possibilities. Both have been addressed by Stalnaker, but without any great emphasis, and the solutions he proposes have not been taken up with the same enthusiasm as the 'big picture'. If I am correct both problems are to be solved in essentially the same way: by investigating the agent's *conscious* beliefs and uncertainties, and then by specifying what should be done with everything that she is not consciously aware of.

In the second half of the chapter I will briefly summarise some work that

has already been done in this direction (not primarily intended for linguistic applications), in the growing literature on awareness for epistemic logic and economic applications. There are two distinct themes in this existing literature, corresponding to different ways of dealing with whatever the agent is not aware of, and these themes provide the division into two parts of the remainder of the dissertation.

To begin with, though, I want to introduce the basic framework that is taken for granted throughout: the possible worlds analysis of meaning, and in particular of belief.

2 · *Possible worlds for belief*

The possible worlds representation of propositional content has a long philosophical history which I do not intend to recap here. I take as my departure point the theory Robert Stalnaker has proposed for belief (in [Sta84]) and conversational context (in papers collected in [Sta99]). To introduce the notions we will focus in this chapter on the simplest case, that of a single believer; the linguistic applications will become more complicated as they include at least two conversational participants.

According to Stalnaker the proper representation of a proposition (the semantic object of a belief attribution or truth judgement) is as a set of possible worlds. A possible world is a "way things might have been", and a proposition can be identified with a set of these worlds; the proposition is true at each world in the set, and false at each world outside the set. While a sentence expresses a proposition (its meaning) via systematic rules of interpretation, the meaning does not need to contain or reflect the structure present in the sentence. (The conjunction of two sentences expresses the proposition that is the set intersection of their meanings; this proposition does not 'remember' that it is a conjunction, it is simply a set of worlds.) If propositions are represented in this way then beliefs can be as well: an agent's epistemic state can be represented by the set of worlds she holds possible; she then believes every proposition that is true in all those worlds (that is, since propositions are simply sets of worlds, every superset of her belief state is a belief she holds; see Figure 1.1).

This shallow description conceals philosophical depths which I do not intend to plumb.[1] Stalnaker's book *Inquiry* [Sta84] is an elaborate argument for and application of this picture; in particular see Chapter 3, "Possible Worlds", for the light metaphysical commitments of taking "ways things might have been" seriously, and Chapter 4, "Belief and Belief Attribution", for the main arguments in favour of representing belief in this way.

[1] Somewhere hidden in those depths is Stalnaker's rejection of David Lewis's "centered worlds" (essentially world-individual pairs). I mention this because Stalnaker's position is perhaps not as popular as Lewis's; so far as I can see either is compatible with everything I will say about unawareness.

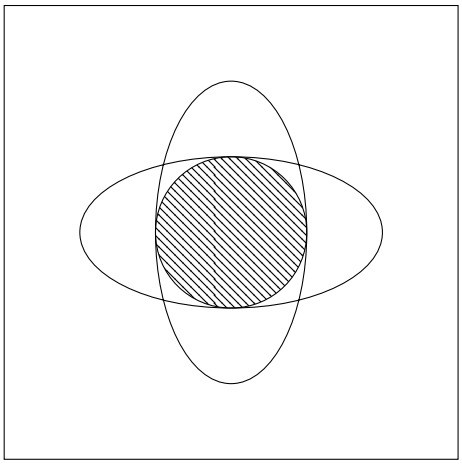


Figure 1.1: A belief set and some beliefs. The square indicates the space of possibilities; the shaded region is a belief set. The two ovals are propositions entailed by the belief set, that is, they are propositions believed by the owner of the belief set.

It is important for Stalnaker's metaphysics that a "way things might have been" is a *full* specification: any question we can think to ask about the state of affairs in some possible world w would be completely answered if we could inspect w in sufficient detail. While this may be metaphysically sound, it is difficult to reconcile with the notion that we use possible worlds to represent beliefs 'in our heads'. There will be a multitude of invisibly fine distinctions that we are unaware of, so that the possibilities we distinguish between should to some degree be underspecified.

Stalnaker gives the obvious answer to this objection: an infinitely finely graduated space of possibilities may be partitioned to various degrees of finegrainedness, by ignoring some differences and paying attention to others.

> [T]here are are surely an infinite number of possible worlds compatible with anyone's belief state. But a believer's representation of a space of possible worlds need not distinguish between them all. Just as a finite perceiver may see a space which consists of an infinite number of points, so a finite believer may represent a space of possible worlds which in fact consists of an infinite number of possible worlds. [Sta84, pg. 69]

We have here a clear distinction between a space of metaphysical possibilities, very large and very detailed, and a representation in the mind of some believer

which is much more limited — a notion which will recur throughout this dissertation. Another recurring notion is that our finite believer *is unaware of* some distinctions which are (potentially; metaphysically) present.

This idea has been given some formal treatment. [Hulo2] noted the connection with partition semantics for questions, and Seth Yalcin devotes a chapter of his recent dissertation [Yalo8] to the idea. Yalcin calls a proposition "accessibly believed" if it makes distinctions that the believer is aware of, and "implicitly believed" otherwise (see Figure 1.2).

As the title of *Inquiry* suggests, Stalnaker is predominantly concerned with *conscious* belief and investigation; similarly, and most naturally, when he applies this theory to a representation of linguistic context it is mainly the representation of explicit and conscious epistemic states that he considers. This is not to say that he ignores implicit or unconscious belief. Indeed, one benefit of a possible worlds representation for belief is that *implicit* beliefs can be given a very natural characterisation: the bus driver who stamped your ticket this morning believes that you are not a disguised lizardman from Mars, not in the sense that he is aware of this possibility and has rejected it but simply because it does not hold true in any of the worlds in his belief state.

However there is a distinction to be made here which Stalnaker does not appear to find particularly important. Some implicit beliefs (that you are not a lizardman from Mars, that Big Ben is larger than Frege's left earlobe) are uncontroversial whether attended to or not. Others, though, seem *unstable* when they are given explicit attention. Here is an example.

Example 1.1: Walt's interview. *Walt hauls himself blearily out of bed at eleven on a Saturday morning. He's staring into (not yet drinking) a cup of coffee when Perky Pat waltzes into the kitchen. "Back so soon from your interview? Did it go well?" she asks. With horror he realises that he is already half an hour late for a job interview on the other side of town.*

No-one should have any difficulty understanding the epistemic condition Walt finds himself in. It may seem strange to refer to his attitude as a 'belief' that he does not have an interview to attend. However it is incontrovertible that he behaves *as if* he believes he has nothing to do that Saturday morning. In that sense we should be willing to say that he has an *implicit* belief: in all the worlds he is actively considering as possibilities he has nothing better to do on a Saturday morning than recover from Friday night. However equally clearly, when this implicit belief is brought to his awareness (or explicit consideration), it is immediately discarded.

There are also cases intermediate between lizardmen and forgotten appointments. Suppose I ask you if you're sure you locked your bicycle (or your car, or your front door) this morning, or whether you turned off the gas after

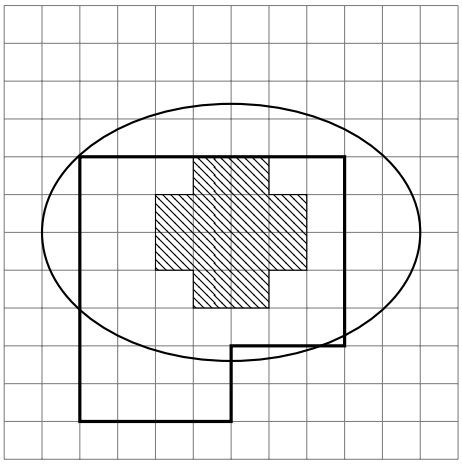


Figure 1.2: Beliefs under finegrainedness restrictions, after [Yal08]. The grid represents the distinctions the agent is capable of making; the shaded area is again a belief set. The oval is a proposition that is *implicitly* believed, the angular proposition is consciously/explicitly/accessibly believed.

cooking dinner last night. If you are somewhat forgetful (as I am) you may have to admit that you're not *certain* that you have done so. If you are habitually forgetful, or merely somewhat paranoid, you may even be prompted by the question to quickly go and check, just to reassure yourself.[2]

All these implicit beliefs have a kind of (intuitively) negative character: they could all be described as beliefs that something abnormal (a job interview, a mistake in a familiar routine, a house-fire) *has not* happened or *is not* happening. Here is another example of such a 'negatively characterised' implicit belief, of a slightly different kind.

Example 1.2: Walt's keys. *Ten minutes later Walt has endured a blistering rebuke and rescheduled his interview. He brushes his chin, shaves his teeth, and hasn't time to sort it out: he'll have to run a few red lights as it is to make the new appointment on time. And where are his keys? Not in his pocket, not on the nail behind the door, not beside the phone, he's searched all the normal places three times. He's staring again*

[2]There is a distinction to be drawn here between practical belief and theoretical certainty. If I ask if you're certain that there isn't a fire starting at this very moment in a far-off corner of the building, you would be displaying evidence of a mild psychological dysfunction if you felt compelled to reassure yourself by checking — this even though your belief that there is no fire is in a sense even less well-grounded than your belief that you have indeed locked your bicycle, even if you cannot remember doing so. Nonetheless, in either case there simply *is* no uncertainty (be it practical or merely theoretical) until attention is directed to the question.

into his coffee, hoping for inspiration, when Pat helps him out for the second time that morning. "Did you leave them in the car when you came in drunk last night?" Walt slaps his forehead (immediately wishing he hadn't) and runs to the car. Alas, the keys are not there either; Pat calls him a taxi (and an idiot), he is late to the interview and doesn't get the job, and later that afternoon she finds the keys in the dishwasher.

Again, the various epistemic states Walt passes through should be familiar, in general form if not (I hope) in specific detail. But if we are to represent them according to Stalnaker's theory we find some interesting difficulties.

The question is where in Walt's epistemic state we should place a world in which the keys are in the car (for this is clearly a "way things might have been"). If it is included as an individuated possibility in his belief set we are unable to explain his behaviour: even if he is not certain if that possibility is the *actual* one, he should certainly act to investigate it.

Perhaps we might combine that world with others in which the keys are *not* in the car, in the same way that an infinite space of worlds can be subdivided into a finite number of distinct possibilities by aggregating sets of worlds whose differences the agent does not attend to. But intuitively this is not the kind of case that prompted Stalnaker to make this suggestion. We might happily aggregate possibilities in which the keys hang from the ignition at slightly different angles —these are distinctions that we can imagine that Walt does not make— but their presence or absence in the car seems such a salient distinction that we should be reluctant to say that Walt does not distinguish the two possibilities from each other.

It seems we must exclude this world from his belief set entirely, but on what grounds? After checking his pockets Walt has good grounds for excluding the world in which they contain his keys from the realms of possibility, but obviously this is not the same kind of case. We can see an immediate difference too in his behaviour after hearing Pat's suggestion: if she had said rather "Did you leave them in your pockets?" he would simply answer "No, I've checked there."

What is missing here is a distinction between two different ways of excluding a world from one's epistemic state: one can have examined it and ruled it out with evidence (as when Walt checks his pockets and establishes that the keys are not there), but one can also have failed to attend to it in the first place. We can fail to attend to entire possibilities as well as to the distinctions between them; failing to attend to *whether p* conflates p-worlds with not-p-worlds, but failing to imagine *the possibility that p* means that only not-p worlds are available to be wondered about.

Just as in the first example, it is strange to claim that Walt 'believes', before Pat's helpful interjection, that the keys are not in the car. However by describing the same state in different terms we improve matters considerably: what he

believes, stated positively, is that the keys *could only be* in his pockets, by the phone, or on their nail beside the door. This belief of course entails the other, but being aware of one need not involve being aware of the other; this is what makes Pat's question (drawing attention to the possibility he is overlooking) helpful.

2.1 · *A failure of attention*

This is an intuitively satisfying solution to describing Walt's predicament. He has failed to consider the car as a possible hiding place for the keys, and what Pat achieves with her question is exactly making him aware of this possibility. As with finegrainedness, Stalnaker has anticipated this possibility, but he doesn't seem to make a firm distinction between 'beliefs' of this kind, held only due to unawareness, and of the more common and reliable sort.

Discussing riddles, he writes:

> More interesting than the case of propositions believed but too obvious to be noticed are those propositions taken for granted only because they are not noticed. With riddles and puzzles as well as with many more serious intellectual problems, often all one needs to see that a certain solution is correct is to think of it—to see it as one of the possibilities. [...] One has beliefs, or presuppositions, which exclude the correct answer. [Sta84, pg. 69]

This seems right for joke riddles such as "What is brown and sticky?"[3] In this case we can even point to the belief at fault: that "sticky" describes the property of sticking to things, rather than that of being similar to a stick.

Perhaps the same kind of description can be applied to Walt's predicament; after all, this is in a sense what the belief that the keys are only in one of the three places he searches comes down to. What about his implicit belief that he has no job interview that Saturday morning though? Here there doesn't seem to be a determinate belief (or presupposition) which *gives rise to* the 'belief' that there is no interview, unless it be simply that belief itself. Stalnaker uses the term "presupposition" (with a technical definition that need not concern us), covering both ordinary (explicit) beliefs and things that we might be less comfortable calling 'belief' without some sort of hedge. Stalnaker, that is, emphasises the similarity of the two cases by choosing a term suitable for both.[4]

[3] A stick. Most native English speakers will have encountered this 'joke' on the school playground; apparently the characterisation of excrement as sticky is less salient for speakers of other languages. I apologise to any readers who through no fault of their own failed to get the joke and now feel left out.

[4] It should be remarked that a Stalnakerian presupposition is primarily a *public* attitude used in explicating conversational behaviour; I am mixing terminology from his investigations of individual belief and of shared conversational context, for the sake of simplicity.

I want instead to emphasise the differences between them. I will use the term ASSUMPTION (introduced in this context in [FJ07]) for a Stalnakerian presupposition (or 'belief', with scare quotes) of this character: one that is held due to lack of awareness of alternative possibilities rather than due to consideration and evidence. In describing Walt's epistemic state we need to distinguish between possibilities that are excluded by assumption (the keys being in the car, or the dishwasher) and those excluded by evidence (the keys in his pocket, by the phone, and so on).[5]

3 · *Assumptions*

Walt is AWARE of, ENTERTAINS, or ATTENDS to certain possibilities, while others he ignores; some of those entertained are then RULED OUT on the basis of evidence.[6] These eliminated possibilities have a different status to those that have never been entertained at all; they are available for examination ("No, I already looked there") in a way that non-entertained possibilities are not.

3.1 · *Sentential awareness*

One key difference between assumptions and 'proper' beliefs is of course whether they are *consciously* held or not. We can imagine asking Walt to describe all his beliefs and uncertainties regarding his plans for Saturday. If he did so he would never mention any interview: not to express his 'belief' (assumption) that he does not have one, and not even to express the tautological belief that he either has an interview or doesn't have one. By contrast we can imagine that if he had had some training in logic (and was not so hung over) he might be led to say "I don't have work today, so I suppose logically speaking I either do or don't have work today."

Suppose we could successfully elicit a complete recital of Walt's conscious beliefs.[7] We can imagine this, idealistically, as a consistent set of sentences in some formal language. The suggestion I want to make is that "interview" is not a term in this language (at least before he has realised his mistake); switching examples, while searching for his keys the term "car" does not appear, while "key" most certainly does. We should think of Walt's lack of awareness of these

[5] We must be careful how we phrase descriptions of assumptions, if we are not to give the wrong impression: saying that "Walt assumes the keys are not in the car" wrongly suggests that this assumption has some distinct status in his epistemic state. In fact it follows only as an entailment from a more general assumption that they are not anywhere but the three places he expects to find them, or equivalently but positively, that they *can be* only in one of the three places he expects them to be.

[6] Again we must distinguish between 'properly' ruling out (the foundation of knowledge, philosophically considered) and 'practically' ruling out (the foundation of ordinary belief).

[7] That this is probably impossible in finite time is not very interesting. More relevant for the current discussion is the practical observation that articulating some of these beliefs may after all prompt Walt to remember the interview, changing his epistemic state through the very act of trying to describe it.

possibilities as restricting the language he has available for *self-ascription of beliefs*.

If we can spell this out in detail, we can keep a possible-worlds representation of belief, while making the distinction between conscious beliefs and assumptions: both are propositions entailed by the worlds held possible according to an agent's epistemic state, but conscious beliefs are describable in his language of belief (self-)ascription while assumptions are not.

Indeed, we can achieve even more than this: the same language will express precisely how finegrained the agent's representation of possibilities is. Two (entertained) possibilities can be distinguished by Walt, according to this story, if he can say what would make one but not the other hold. His language of conscious belief establishes which possibilities can be distinguished, while his unconscious assumptions exclude some from consideration entirely.

Suppose Walt is not attending to the question whether p. This might be for either of two quite different reasons: he might be *indifferent* between p and not-p alternatives (perhaps in the normal sense of the word, or perhaps because he simply is not conceptualising the difference between p and not-p), or he might assume that the question is already *settled* (that is, he either assumes that p, or that $\neg p$). The result of calling attention to p in both cases is the same: he becomes aware of both p worlds and not-p worlds as distinct 'live' possibilities (the conscious belief attitude he holds to them —whether he gives them any credence— is a separate matter). However the two mechanisms look different: in the first case a more finegrained distinction is made between possibilities that were previously considered equivalent, while in the second case a genuinely new possibility is brought into the light of conscious consideration.

It may be quite difficult to observe (at least with any certainty) the first kind of awareness dynamics. Having the capacity to make a distinction, after all, is no promise that the distinction will in fact be reflected in any observable behaviour. The second kind of dynamics, however, are often extremely visible; Walt's forehead-slap is the typical sign of an assumption being OVERTURNED.

3.2 · *Overturning assumptions*

Why is mentioning the car enough to overturn Walt's assumption that the keys are not in it? This effect shows a second important distinction between assumptions and conscious beliefs: consciously held beliefs are typically justified by various kinds of evidence, while assumptions need not be. If you (consciously) believe P and are confronted with evidence that P is not after all the case, you face a difficult task. Revising your belief that P will likely also require changes to a wide range of attendant beliefs, especially those that partly justified or were justified by the belief that P. This difficulty has been recognised in the formal literature on belief REVISION [AGM85]. Revision is contrasted with monotonic belief UPDATE, in which new information is learned that is consistent with the

previous epistemic state; belief update is easily represented in a possible worlds semantics, simply as the elimination of worlds, but belief revision requires a much more complex operation.

The overturning of an assumption has something of an intermediate character: it resembles belief revision in a formal sense (a previously held conviction is overturned, which cannot be represented by elimination of possibilities) but intuitively it is much more similar to belief update. The key property here is the ease with which an assumption is let go: belief revision involves a rearrangement of a whole network of attendant beliefs, while assumptions seem to be overturned more or less in isolation.

This is to be expected, once we recognise what kind of 'belief' an assumption is. A belief is formed consciously by eliminating possibilities according to evidence; if that belief is to be overturned then the evidence that lead to those eliminations must also be reexamined. But an assumption is not based on evidence at all: it is based, in fact, precisely on *not* considering all the available evidence. Part of the difficulty of belief revision is the attitude to propositions that are not direct consequences of the revised belief but are also not independent: those that provide support or evidence for a proposition now deemed false, and that therefore must be considered suspect. These simply are not present in the case of an assumption, which makes the revision process much simpler.[8]

If assumptions are not justified by evidence, though, where do they come from? Intuitively it is clear that not just any proposition is a plausible candidate to be assumed. If the explanation of Walt's epistemic changes had rested on his assuming that the keys were in the dishwasher, then either the explanation or the use of the term "assumption" would have to be questioned. But what makes "The keys are not in the car" a plausible assumption while "The keys are in the dishwasher" is not?

3.3 · *Where assumptions come from*

Assumptions come fundamentally from a failure of imagination. We are aware of fewer possibilities than in fact exist, and whatever we mistakenly ignore we have assumed away. It might seem from this description that we can easily 'read off' a specification of assumptions from checking what the agent does not attend to: Walt does not attend to the possibility of his interview, and so assumes it does not exist; he does not attend to the possibility that his keys might be in the car, and so assumes that they are not.

What makes Walt's assumptions deceptively simple-looking is that they

[8] Formally speaking we have to *add* worlds but never to remove any that we were entertaining. After an update by belief revision proper the new set of worlds may be disjoint from the old one, and it is the complicated relation between these disjoint sets that makes this process difficult to describe formally.

have a clearly negative character: there is some particular thing he is not thinking about (an interview; the car), and his assumption roughly corresponds to the absence of that thing (more about this in a moment). But there are plenty of assumptions that do not have this negative character. Every day when I arrive at work I assume that my key will open the door of my office; I likewise assume that the same key will not open any other office in the same hallway. I assume that my computer is still on my desk and running, and that I can log on to do my work. These are all assumptions in the technical sense I intend: they are implicit beliefs that I act on without any conscious attention to the concepts involved or to whether the beliefs are reasonable or not.

The existence of these 'positive assumptions' means that we cannot simply read off assumptions from unawareness: just because I do not attend to the question of whether this particular key will work in this particular lock, you cannot immediately tell whether I assume it will (the lock of my office) or it won't (someone else's office).

Still there seems to be a distinction between these cases and Walt's assumption that his keys are not in the car. We might say that the former are assumptions due to *induction*, while the latter assumption is due to *limited awareness of objects*.

3.3.1 · *Awareness of objects*

Walt's situation, after looking for the keys in every place he can think of, is quite likely familiar to the reader, who will recognise the feeling that accompanies it: a nagging sense that if one could just remember a few more places the keys *might* be, one would immediately know where they *are*. The things being ignored here are not primarily propositions, cognitively speaking, but *objects*: possible hiding places for the keys.

This distinction need not, per se, be represented in our formal theory. Stalnaker has argued [Sta84, p. 61] that 'aboutness' does not need to be encoded in the structure of a proposition: the fact that the sentence "Socrates is mortal" is about Socrates does not need to be explicitly represented in the structure of its meaning; the set of worlds in which Socrates in fact *is* mortal (the proposition, in other words) will do just fine.[9] In the same way, we can use the possible worlds machinery to represent attending to or ignoring an object: the 'aboutness' of that attention need be nothing more than a constraint on which worlds are entertained and which assumed away. Nonetheless, in this case we cannot entirely ignore the cognitive intuition, for we need to define our agent's language of self-ascription of belief: it is here that the absence of a term "car" appears.

[9] Aboutness is not done away with completely, of course. That "Socrates is mortal" is about Socrates is encoded in the manner by which we attach exactly that proposition to exactly that sentence. What is given up is the idea that this relation is directly visible in the meaning of the sentence itself.

If "car" is not in Walt's conscious vocabulary of awareness, then he attends to no propositions 'about' the car. But we must still specify which beliefs he holds, if only implicitly, about these propositions. That is, we must justify the absence of any worlds in which the keys are in the car from his assumption set. Here the criterion seems to be relevance: whatever objects Walt isn't aware of, he assumes to be (and thus to have properties making them) *irrelevant* to his problem of finding the keys. It is this that justifies excluding the world where the keys are in the car from his conscious epistemic state, just as it justifies excluding the world where the keys are in the sugar jar: by implicitly judging the sugar jar and the car irrelevant (by not attending to them) Walt has ensured that he will not consider the possibility that they are hiding places for the key.

One might reasonably ask what is cause and what effect here: does Walt ignore the car because he considers it irrelevant, or does he consider it irrelevant because he has not attended to it properly? In fact neither perspective seems quite right; the connection between irrelevance and unawareness has more the nature of a consistency constraint than a causal relation.

What about other properties of the car, such as its colour (likely irrelevant for the key question) or even its existence, or at least its presence in the garage (which in fact is quite pertinent)? Assumptions about these properties, I suggest, are settled in the same way that the assumption that your bicycle is locked gets formed: by *induction*.

3.3.2 · *Inductive assumptions*

A large number of assumptions do not seem to derive from ignoring objects. Instead, they are 'properly propositional' in the sense that the natural way to express what is being ignored is "the possibility that (some proposition) p does (or does not) hold". These seem to come from various kinds of inductive reasoning, which neatly explains our intuitions about which assumptions are plausible and which are not.[10]

There are assumptions that are clearly induction from previously observed instances; the assumption that the building you are currently inside is not on fire, for instance, that the company you work for has not gone out of business overnight, and so on. (It is important for these examples to recall that 'assumption', in the sense I use the term, refers only to beliefs held due to *inattention* and *unawareness*. In both these cases we might continue to believe the proposition under consideration even after attending to it, but we should have to admit that we had no grounds but those of precedent to do so. These are 'assumptions' in the normal sense but not in the technical sense I intend.)

The assumption that "sticky" refers to the property of sticking to things

[10] "Inductive reasoning" is perhaps too strong: assumptions take in 'inductive' generalisation from a single observed instance, which hardly deserves to be called reasoning at all. The 'inductive assumption' is that what the agent has seen is everything there is to see.

might come under this heading also, or it might represent a more general tendency we have to resolve ambiguities by fiat rather than with full attention to the range of possibilities. This is especially visible in our attitude to language. Most sentences are strictly speaking ambiguous in some way or another, but in most natural contexts one reading is strongly preferred (this is not a coincidence: if we are to communicate effectively we have to use expressions that can be easily disambiguated to do it with). The observation linking this to awareness is that typically this potential ambiguity does not reach the level of consciousness, not even as a recognition that an alternative potential interpretation is being rejected. In the case of the riddle this resolution is probably based on two distinct mechanisms: that the interpretation being assumed is the more common one, and that it gives rise to a plausible (if scatological) answer to the question and thus doesn't trigger any extra effort towards exploring more unlikely possibilities.

Another kind of inductive assumption might be called "There it's like here": the assumption that conditions are similar in distant places or times (apart from whatever distinctions are being consciously attended to, of course). Inexperienced travellers are constantly surprised by the variations in standards of politeness around the world; actors in period dramas are tall, unmarked by disease, and clearly bathe regularly, and we don't notice any incongruity.[11]

I will not have much to say about the dynamic process of forming assumptions; whether approached philosophically or psychologically the formation of inductive generalisations is a vexed question that I will sidestep as far as possible. But one potential source of assumptions seems clear: *any belief at all* can be converted into an assumption if it goes unchallenged and unexamined for long enough. Having eliminated a possibility (by examination and evidence) we do not in fact cling to it for the rest of our lives; in time it fades from our consciousness and what was once a conscious belief becomes nothing more than an assumption.

3.4 · *What are assumptions for?*

We notice assumptions most often when they are overturned, or when they turn out to be false. Remember, though, that a finite epistemic state supports an infinite number of assumptions; the vast majority of these will never be attended to and would be quickly ratified, rather than overturned, even if

[11] Two days after I wrote this Brendan Adkins asked on his blog "Not Falling Down",

> Why do the same people who complain about sound in space, the proper rigging for catapults, or the relative strength of a katana versus a broadsword never mention the way women in medieval or even Victorian settings are always depicted with shaved legs?

Inattentiveness and assumption is not the only possible answer to the question, but it is a plausible one.

they were. (Big Ben is larger than Frege's left earlobe. My feet have toes, my hands have fingers. I am not a butterfly dreaming it is a PhD student.) In this dissertation, however, we will be particularly interested in the cases where assumptions go wrong: it is these that provide the interesting pragmatic and semantic possibilities that make this notion so important for describing and explaining the various ways we use language. Looking at these cases one might start to wonder whether having assumptions is a smart idea after all. Might we be better off if we could do without them? Life would certainly get a lot simpler, wouldn't it?

The first partial answer is of course that we cannot do without inattentiveness, since the space of possibility is too large and our minds too limited to comprehend it all. However, besides being a negative and rather trivial statement, this does not mean we need assumptions: we might get along fine just using finegrainedness (as in [Yalo8]) to carve the world up into manageable chunks.

A more positive answer is that assumptions make reasoning much easier. In deciding how best to fetch that banana we need not consider the possibility that it will fly out of reach when we get close to it; given the unlikelihood of the possibility and the fact that if it did eventuate we would have no sensible strategy for dealing with the problem, not having to consider it seems a mercy. In other words, assumptions do not just reduce the space of possibility so that it can fit inside a believer's head; they select the possibilities that are most relevant for the problem-solving the believer will have to perform, those that are likely enough to be worth attending to, and also those that the believer can, with sufficient planning, deal with if they eventuate.

If this is what assumptions are for, then, we can distinguish two properties they need to have if they are going to do their job properly. They need to be *generally true*, and they need to be *defeasible*. (These two properties, in turn, suggest why assumptions can typically be described as inductive generalisations.)

That assumptions should be generally true should by now be obvious. Excluding the correct answer by assumption, as in the case of riddles, only makes life difficult. It is impossible to reason outside one's own assumptions (if you are reasoning about the possibility that p, you are aware of it). This means that if we are ever confronted with a genuinely unentertained possibility, we will have to very quickly construct a plan to deal with it. Clearly it will be more comfortable to avoid this kind of last-minute plan formation as much as possible: we should hope that most of our assumptions turn out to be acceptable, most of the time.

And indeed, evolution seems to have equipped us with a magnificently effective machine for forming just the right inductive judgements in just the

right ways, so that they tend to be true. In fact it can be argued that *all* of our beliefs rest upon a foundation of assumption rather than of evidence strictly conceived.[12] The fact that our beliefs tend in the main to be true ones speaks strongly for the reliability of our assumption-formation apparatus.

The other requirement I stated is that assumptions should be defeasible. This is strictly speaking a requirement not on the notion of assumption but on how they are recorded and represented in the heads of believers; they will by their very nature be defeasible, but they need to be represented in a way that makes it easy to revoke them. This is both an observation (assumptions *are* easily revoked, as I argued when introducing the notion) and a prediction: assumptions *must* be revokable with a minimum of difficulty, when contradicted by evidence. This is because they may be confounded by the world; if this occurs the believer should give up assumptions before giving up belief in more tangible evidence. Walt might check twice that the keys are not hanging on the nail behind the door, but if he repeatedly gives up the belief that the evidence of his eyes is correct, rather than the assumption that the keys cannot be somewhere else, we would be justified in concluding that he suffers some psychological disturbance.

Attention or awareness?

I have used "attention" and "awareness" more or less interchangeably, and will continue to do so in the rest of the dissertation. They carry different associations which allow for slightly smoother exposition: "being aware" is a passive state while "attending to" is an active action, and (as we will see) many linguistic

[12] Wittgenstein advances a similar position in *On Certainty*, with the added observation that certain of these assumptions cannot even meaningfully be questioned, without undermining the sense of the language in which the questioning is attempted. Or rather, perhaps, it is not possible to question *all of these assumptions at once*: some basis of certainty is required, otherwise a doubt cannot even be meaningfully expressed.

> 79. That I am a man and not a woman can be verified, but if I were to say I was a woman, and then tried to explain the error by saying I hadn't checked the statement, the explanation would not be accepted.
>
> 80. The *truth* of my statements is the test of my *understanding* of these statements.
>
> 81. That is to say: if I make certain false statements, it becomes uncertain whether I understand them.
>
> 83. The *truth* of certain empirical propositions belongs to our frame of reference.
>
> 96. It might be imagined that some propositions, of the form of empirical propositions, were hardened and functioned as channels for such empirical propositions as were not hardened but fluid; and that this relation altered with time, in that fluid propositions hardened, and hard ones became fluid.
>
> 341. That is to say, the *questions* that we raise and our *doubts* depend on the fact that some propositions are exempt from doubt, as it were like hinges on which those turn. [Wit69, §§79–81,83,96,341]

utterances are best thought of as drawing attention in this active sense.

"Unawareness" can also denote a lack of conceptual grasp; the man on the street is unaware of the principles of higher mathematics, and Darwin was unaware of the existence of DNA when he proposed his theory of natural selection. The examples I am mainly interested are more humble: they concern concepts that the agent in principle understands, but has not thought to apply to the case at hand. "Inattention" is perhaps a more appropriate term for these moments of forgetfulness and lack of insight.

However there is at least one potential confusion risked by using "attention" in this way, which needs an explicit comment. In linguistics attention is often a *relative* quantity, allied with notions such as salience. When the 'focus of attention' is directed at something in particular, it is *withdrawn* from whatever else might have been under discussion. This is emphatically not the sense in which I intend the word. I have tried to use "awareness" particularly in cases where this interpretation seems tempting, since 'becoming aware' is much less seductively misleading than 'shifting attention elsewhere'. I have of course also used "awareness" where it is a technical term in existing literature, as in the three models discussed in the rest of this chapter.

4 · *Three models of awareness*

We have seen how thinking about formal semantics and philosophy of language can lead us to notions of awareness and assumption. There is a growing field of research investigating the notion of awareness in epistemic logic (and related fields such as certain branches of economics), without any attention to language at all; this field has so far largely ignored the notion of 'assumption', which is so crucial for understanding the behaviour of agents suffering from unawareness.

In the rest of this chapter I will survey two very influential systems based on very different notions of unawareness (the logic of general awareness of [FH88] and the subjective state-space approach of [HMS06]), paying particular attention to the space they leave for a representation of assumption. A third proposal that is gathering support is the object-based model of [BC07], which to some extent bridges the gap between the first two models; in particular, while the object-based semantics is a closer formal cousin to the logic of general awareness, the natural notion of assumption it suggests is more closely connected to the state-space approach. These three papers inform the structure of the rest of the dissertation: very broadly speaking, Part I extends models along the lines of [FH88], while Part II combines the object-based approach of [BC07] with the subjective state-spaces of [HMS06].

4.1 · *The logic of general awareness*

Fagin and Halpern's paper [FH88] is generally credited as the origin of the current research field of unawareness in epistemic logic. The paper deals with

the problem of LOGICAL OMNISCIENCE: the property of standard possible-worlds analyses of knowledge that an agent knows all logical consequences of her knowledge. This leads to unintuitive results when we match our models against the cognitive limitations real reasoners suffer; for instance the naive prediction from such a model is that a reasoner should know all logical tautologies and mathematical truths.

Fagin and Halpern point out that 'logical omniscience' is best thought of as an umbrella term covering a number of distinct problems, and they propose several systems dealing with different aspects of the problem. The one we are concerned with here is the LOGIC OF GENERAL AWARENESS (for this chapter the LGA, found in their Section 5).

The essence of this system is a distinction between IMPLICIT and EXPLICIT belief. 'Implicit' belief is a new name for the familiar notion of belief interpreted on Kripke structures: a box modality on a serial, transitive and Euclidean accessibility relation (giving rise to the logic KD45). *Explicit* belief, on the other hand, is modelled via awareness. An AWARENESS FUNCTION for each agent assigns to each world the set of sentences the agent is aware of (note the syntactic nature of this component). The agent explicitly believes a formula φ if she implicitly believes the proposition that φ expresses (standard belief in Kripke structures) *and* she is aware of the formula. (Among other things this means that every explicit belief is also implicit. While formally convenient this grates somewhat on the natural usage of the terms: typically we would say "implicit belief" to mean precisely a belief that is *not* explicit. I will follow the 'logician's terminology' of Fagin and Halpern here, and in similar cases that emerge in the later models, and distinguish beliefs that are not explicit by calling them 'purely' or 'strictly' implicit when the distinction is important.)

The representation of awareness is left completely unconstrained: the awareness function can select any set of sentences whatsoever at each world. (It is not required, for instance, that the agent be aware of $p \wedge q$ whenever she is aware of $q \wedge p$.) For Fagin and Halpern this is an advantage, since it gives their system the flexibility to represent many different kinds of 'unawareness' (in a rather broad sense) leading to different kinds of failures of logical omniscience. For our purposes though it seems we want to interpret awareness as something like *linguistic resource*: that the agent has the conceptual vocabulary to describe explicit beliefs is what distinguishes them from those that are strictly implicit. In particular, Fagin and Halpern discuss a set of constraints on the awareness function that reduces awareness of sentences to awareness of atomic formulae (primitive propositions): the agent is aware of a complex formula iff she is aware of all the atomic formulae that occur within it.[13] All the models I propose

[13] They also treat possible interactions between belief and awareness; for instance, an 'awareness introspection' property that the agent knows what she is aware of can be guaranteed by requiring

in this dissertation have this 'combinatorial' property.[14]

It seems that the LGA provides a fairly close fit for the notion of awareness we need. It is less successful, unfortunately, in representing assumptions.

4.1.1 · *Assumptions in the* LGA

It might seem that implicit belief is precisely what is intended by the intuitive notion of assumption (an assumption is distinguished from a 'real' belief by being strictly implicit while the latter is explicit: the purely syntactic distinction of [FH88]). However there is a further *semantic* distinction that can be drawn, which becomes important for representing changes in awareness over time: the distinction between worlds that the agent 'has in mind' and those that she does not.

In 1957 the BBC current affairs program *Panorama* reported on a bumper spaghetti harvest in Switzerland.[15] Viewers doubtless formed the conscious, explicit belief that the mild winter had been good for the growth of spaghetti, since that was what the programme announced. That this belief rests on the assumption (entirely implicit) that the BBC reports only the truth (the broadcast was of course an April Fools' joke) does not make the belief itself any less explicit.

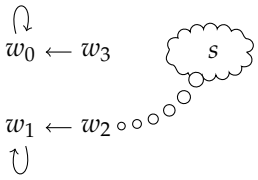
Suppose we were to represent this scenario in a model for the LGA. The model representing a viewer named Vera is shown in Figure 1.3: she (strictly) implicitly believes the BBC is entirely trustworthy ($\neg j$, where j stands for "joking"), and explicitly believes that spaghetti is grown in the south of Switzerland and dried in the alpine sun (s).

The problem with this picture is that it does not distinguish between Vera's attitude, at the actual world w_2, to w_0 and to w_2 itself. Both are excluded from her possibility set, but intuitively for very different reasons: w_2 is a possibility she is *not even imagining* (despite its being actual), while w_0 is a possibility she imagines but has ruled out (on the basis of the BBC broadcast). This intuitive difference shows in a difference in behaviour, once we take changes in awareness into account. If we should politely draw Vera's attention to the date, raising the possibility that the BBC is having a bit of fun, the worlds w_0 and w_2 do not behave the same way in her resulting epistemic upheaval. We expect her to come to hold w_2 possible exactly because it was previously excluded by an assumption; she should *not* come to hold w_0 possible, because it was excluded

that if w' is accessible from w then the awareness function returns the same set at each of the two worlds.

[14] Under this constraint we can think of the agent's awareness as providing a partition on logical space representing a level of finegrainedness, as in Figure 1.2 on pg. 5. This is not in general possible for structures of the LGA; for instance the agent might explicitly believe both φ and ψ but only implicitly believe $\varphi \wedge \psi$, which cannot be represented via such a partition.

[15] The segment can be viewed online at the BBC website: http://news.bbc.co.uk/onthisday/hi/dates/stories/april/1/newsid_2819000/2819261.stm



	j	s
w_0	0	0
w_1	0	1
w_2	1	0
w_3	1	1

Figure 1.3: Vera, a victim of a BBC April Fools' hoax in the 1950's, modelled using the logic of general awareness. Proposition letters j and s stand respectively for "joking" (that the BBC broadcast was not in earnest) and "spaghetti" (that spaghetti is grown on trees). Arrows represent her accessibility relation; the actual world is w_2 and the 'thought balloon' represents her awareness function (shown only at that world): she is aware of s but not of j.

by conscious consideration of the evidence.[16]

The tactic I will employ in the following chapters is effectively to define *two* accessibility relations: one representing the possibilities that the agent 'has in mind' (regardless of her attitude of belief or scepticism towards them) and one representing her beliefs (regardless of whether these are implicit or explicit). Any world she does not 'have in mind' is automatically excluded from her belief set; the worlds she 'has in mind' define her assumptions, in the sense that Vera's belief that there are spaghetti trees in Switzerland is based on her assumption that the BBC broadcasts only factual reporting.

Segue: formal epistemic economics

The approach of [FH88] takes implicit belief and awareness as primitive notions, and then derives explicit belief from them. An opposing theme in the formal epistemic economics community[17] is to take *knowledge* as primitive (roughly corresponding to explicit belief, although as we will see the identification of the two is not without problems), and defining (un)awareness as a derived notion.

[16] This point is liable to misinterpretation. I do not mean that the BBC's announcement should be treated as *true* (which would certainly prevent Vera from reinstating w_0, but would also keep w_2 out of the picture). The broadcast itself remains as an objective fact, and that fact is incompatible with the *combination* of there being no spaghetti trees in Switzerland and the broadcast being truthful. This combination is present in w_0, and thus ruled out by the objective evidence (the existence of the broadcast), even after an awareness change.

[17] This is the best term I could find for the sub-field of economics exemplified by the biannual conference TARK (Theoretical Aspects of Rationality and Knowledge) and results such as the no-agreeing-to-disagree and no-trade theorems [Aum76; MS82]. The field stands in roughly the same relation to general economics as model-theoretic semantics does to general linguistics. I don't imagine that all "economists" are concerned with common knowledge any more than all "linguists" make use of Kripke structures, but I will continue to use the terms as though this were the case, since these are the only linguists and economists I am directly concerned with.

The idea begins with the observation that unawareness leads to a failure of negative introspection: if the agent is unaware of p, she can fail to know p but not know *that* she does not know it. Conversely, she is aware of p just if she either knows that p or she knows that she does not know it. S5 knowledge of course supports negative introspection; the standard approach in economics to modelling agents without negative introspection has been the use of NON-PARTITIONAL INFORMATION STRUCTURES: Kripke structures for S4. Unfortunately, [MR94] proved that such structures cannot be suitable for modelling unawareness: adding a symmetry axiom requiring that the agent be aware of φ just if she is aware of $\neg\varphi$ (eminently reasonable under our 'conceptual vocabulary' interpretation of awareness) again yields S5. Modica and Rustichini also investigated a larger class of models where knowledge is given not by an accessibility relation but by a function mapping each event[18] $E \subseteq W$ to the event of knowing E. While such models can incorporate symmetry without collapsing to S5, Modica and Rustichini showed that they nonetheless only give rise to trivial unawareness (either the agent is aware of everything or she is aware of nothing).

This negative result has acquired the label "Standard state-spaces preclude unawareness", after a later paper extending the treatment [DLR98]. This later paper gave explicit attention to a major problem with the knowledge-based analysis of awareness: the treatment of tautologies. Standard models based on possible worlds translate any tautologous sentence to the (one and only) necessary proposition (the entire state space). But on the idea of awareness as representing conceptual vocabulary, the agent might very well be aware of $p \vee \neg p$ while remaining unaware of $q \vee \neg q$. For models with syntactic awareness as a primitive this is of course no problem, but economists seem to have viewed such models with some scepticism because of the 'pollution' of semantics with syntax.

Instead, the economics community has turned its attention towards non-standard state spaces, in particular towards models in which not every 'world' represents a genuine fully-specified possibility. If a proposition is a set of *partial* possibilities, rather than full possible worlds, than a partial possibility unspecified for the value of p might reasonably fall outside the proposition expressed by $p \vee \neg p$, thus reinstating the ability to distinguish between even tautological sentences based on the vocabulary employed. The next model we will consider exemplifies the type.

4.2 · *The subjective state-space approach*

[HMS06] introduces a model (which I will call the HMS model) in which some states are associated with *partial*, rather than total, valuations of the set of atomic

[18] An "event" to an economist is a "proposition" to a linguist: a subset of the set of worlds or possibilities.

formulae. The state space is constructed from a lattice of disjoint subspaces, where each subspace intuitively corresponds to the language generated by a particular vocabulary.[19] The topmost subspace contains 'real possibilities', or full valuations; worlds in lower subspaces are *partial* valuations. An agent unaware of p 'sees' only worlds in a subspace whose vocabulary does not include p, thus whose valuations do not assign a truth value to p (the agent is unaware of p at w_4 in Figure 1.4, for example).

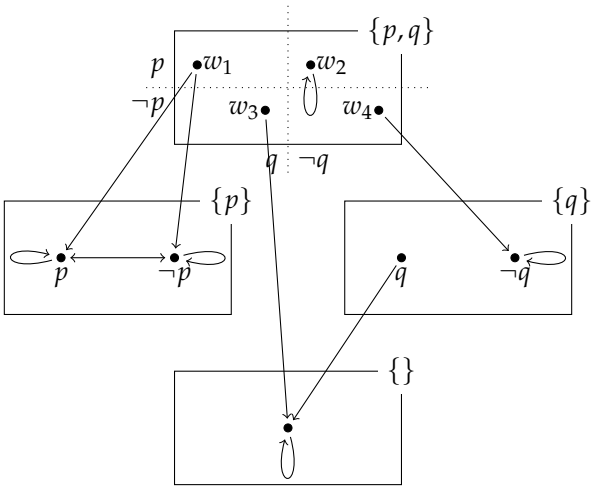


Figure 1.4: An example of a simple single-agent HMS model. Arrows indicate the agent's accessibility relation; her awareness at each world is given by the vocabulary of the subspace she sees into at that world. At w_1 she aware of p but not q, and uncertain whether p or $\neg p$; at w_2 she is aware of both p and q and knows exactly which holds; at w_3 she is aware of neither; and at w_4 she is aware only of q and knows that $\neg q$ holds.

The subspaces are partially ordered according to the richness of their vocabularies, and there are projection functions (shown in Figure 1.5 overleaf) saying how a world in a high space 'appears' when viewed according to the limited vocabulary of a lower space. (Typically several worlds in any given higher space will project to a single world in a lower space; for instance two worlds differing only in the valuation of p would project to the same world in a space where p was not part of the vocabulary. Each world projects to only one world

[19] The construction in fact given in the paper makes no formal reference to vocabularies and so on; I follow the presentation of [HMS08]. 'The' HMS model is considerably more protean than I pretend here; see especially note 22 on pg. 26 for a more recent variation with quite different properties.

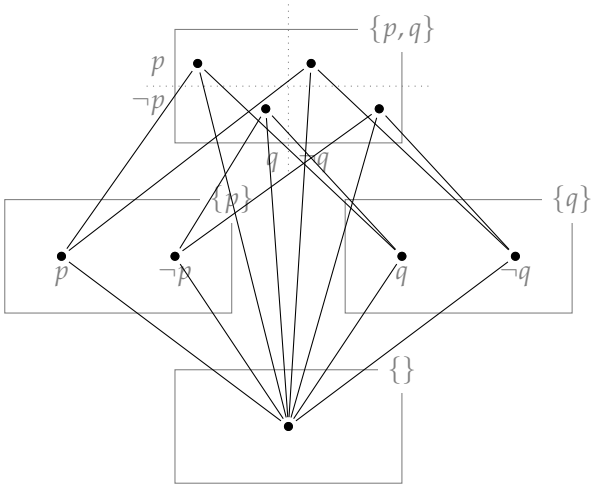


Figure 1.5: Projection relations between worlds. Each world projects (downward) to exactly one world in each lower subspace; it will project (upward) to several in a higher space. Downward projection represents "how the world appears" in the vocabulary of a lower subspace; the world where both p and q are true appears as a world where p is true in the subspace with vocabulary $\{p\}$, and where q is true in the subspace with vocabulary $\{q\}$.

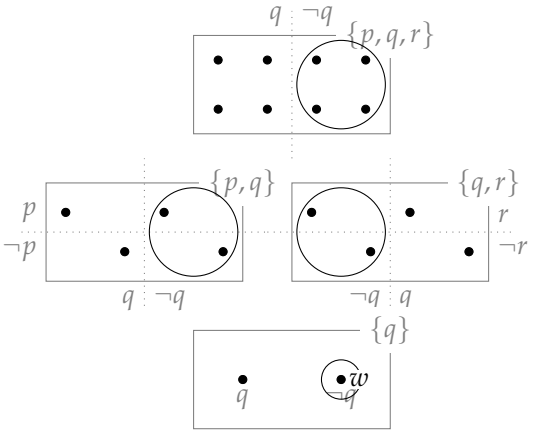


Figure 1.6: Part of an HMS model showing an upward cone (several subspaces are omitted). The circles all project down to the world labelled w; the event generated by the set $\{w\}$ is the upward cone that is the union of the circles.

in each subspace, however, and a world and its projection always agree on the valuation of whatever vocabulary they have in common.)

The lattice structure of the subspaces constrains the accessibility relation: from w an agent can only see downward in the structure (to spaces with less expressive power) or 'across' to worlds within the same subspace as w. Further constraints generalise reflexivity, transitivity, and so on to the lattice structure: if w sees v in a lower subspace, for example, v cannot see w (because this would be looking 'up' the lattice into a space with higher expressive power) but the analogue of reflexivity requires that v see *the projection of* w into the same space v inhabits.

Instead of just being a set of worlds, an event (or proposition) in this setting is an UPWARD CONE through the lattice structure, generated from a single subspace: for some set of states B lying in a single subspace it contains all states that project into B (that this is an upward cone follows from a requirement that projections commute down the lattice of subspaces). See Figure 1.6 on the facing page for an example.

Among other things this means that the negation of an event is not simply its set-theoretic complement: such a set would typically not be an event. Instead we generate an upward cone from the complement of B *in its subspace*; this process ensures that an agent aware of the event P (one who sees the subspace generating P) will also be aware of its negation (the symmetry constraint required by [MR94]).

Awareness is defined based on knowledge (the agent is aware of P if she knows P or knows that she does not know P), and knowledge is given a standard definition: she knows that P at w (where P is an event, i.e., an upward cone) if all her worlds accessible from w lie within P. However the structure of the subspaces, and the fact that no agent can see 'upwards' in that structure, ensures that she only knows that P from worlds that have enough vocabulary to describe P, that is, from worlds in the same subspace that generates P or from higher in the lattice.

At first glance this looks a lot like supervaluations. Worlds in subspaces are partial models, and the projection relation tells us which more complete models they may 'grow into'. However there is an important difference, to do with the very notion of unawareness. Under a supervaluational definition of knowledge, the agent always knows a tautology such as $p \vee \neg p$ (since it is true at all supervaluations where it is defined); exactly this has to be avoided in a model of awareness. One way to see the distinction is to take the language-oriented definition of knowledge (see Figure 1.7 overleaf). In a supervaluational story, the agent knows φ iff φ holds everywhere in the projection of her belief set on the highest subspace (that is, in all supervaluations of the partial valuations that make up her belief set). In the HMS model, she knows φ iff φ holds everywhere

in projections of her belief set on *any* (weakly) higher subspace, including the subspace where her belief set itself lives. "Holds everywhere" means *is defined and* holds everywhere; this makes no difference for the supervaluational version, since in the highest subspace all sentences of the language have definite truth-values, but it matters a lot for the HMS model: a tautology such as $p \vee \neg p$ is only defined in the agent's belief set if she is aware of the proposition letter p.

This has enormous consequences for the representation of assumptions.

4.2.1 · *Assumptions in the* HMS *model*

In fact it means that non-trivial assumptions are entirely ruled out. Suppose the agent is unaware of q, and sees only worlds in a subspace with vocabulary $\{p\}$. She holds possible only the world where p in fact holds, and she thus knows that p. If we follow the projection relations backwards from her belief set we can collect the sets of *fully specified* worlds (in the highest subspace, with most extensive vocabulary) that project to each of the worlds she holds possible. Whatever knowledge she can acquire by eliminating worlds in the subspace whose vocabulary she can use, it must deal atomically with these 'knowledge units': eliminating a world in the lower space eliminates its entire inverse projection set, so no knowledge generated from the lower subspace can ever 'cut across' such a set (see Figure 1.8 on the facing page).[20]

But this is precisely what we want assumptions to do! Every world in the subspace with vocabulary $\{p\}$ is the projection of two worlds in the higher space: one where q holds, and one where $\neg q$ does. Think back to Vera and the spaghetti: she is unaware of j and believes s. To represent her assumption that $\neg j$, we need to *separate* the two worlds that project to the (partial) world she imagines: the one where $\neg j$ holds belongs in her belief set,[21] while the one where j holds does not.

Here is another way of looking at the problem. The set of worlds generating Vera's knowledge is generated as an upward cone through the lattice of subspaces, so it includes some 'worlds' that are more partial than others. Her knowledge is whatever holds at *all* these worlds, and the construction restricts this in two distinct ways. First there is the vocabulary restriction: since some worlds are partial and make no mention of j, she cannot know that $j \vee \neg j$ (the observation of [DLR98], that unawareness even affects tautologies). We might be tempted to say that she knows this *implicitly*, however, since it is nowhere contradicted in this set of worlds; it is certainly true everywhere in the *complete*

[20] Note again that this applies only to the model with a language as presented in [HMS08]; see note 22 on pg. 26.

[21] I used the word "belief" deliberately instead of knowledge. The upward cone construction can be seen as a conservative way of ensuring the factivity of knowledge under unawareness: the agent is prevented from knowing *anything* about whatever she is unaware of.

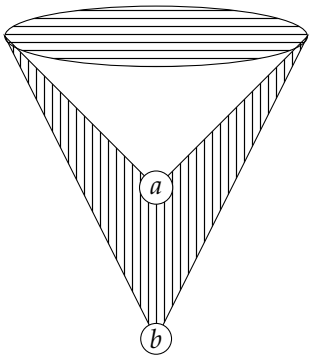


Figure 1.7: Supervaluations compared to HMS structures (in abstract visualisation). The cone from a represents "$p \vee \neg p$", generated from the subspace where only p is defined; the cone from b is generated from the belief set of an agent unaware of p, thus from a lower subspace. Supervaluations quantify only over the worlds in the horizontally shaded region; HMS knowledge quantifies over the entire diagram, including the vertically shaded region where p (and thus $p \vee \neg p$) is undefined.

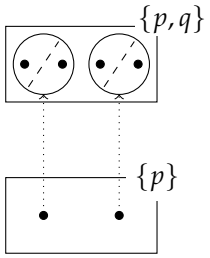


Figure 1.8: HMS models cannot represent assumptions. Dotted arrows are upwards projections. If the agent is unaware of q (that is, her set of accessible worlds lives in the lower subspace) then it treats the circles in the upper subspace atomically: each circle is either entirely inside or entirely outside the upward cone generated by her accessible worlds. Her beliefs can never 'cut across' the dashed lines.

worlds in the set, which intuitively correspond to the real possibilities still left open. She cannot know that j, though, for another reason as well: there are worlds in that set where it *is* defined but does *not* hold. If an assumption corresponds to some kind of implicit belief, it will not be found in this set.

4.3 · *Object-based unawareness*

The object-based model of [BC07] (OBU, for "object-based unawareness") is a *first-order* model, in which the agent's awareness of sentences is generated by her awareness of objects. It bridges the gap in various ways between the logical approach of [FH88] and the economic approach of [HMS06].

An event in OBU semantics is a pair: a set of worlds (the "sense" of the event) and a set of objects (the "reference"). The reference contains the objects the event is 'about'; the tautologies $P(a) \vee \neg P(a)$ and $P(b) \vee \neg P(b)$ have the same sense but different references, as you would expect. The model contains an accessibility relation on worlds but also an awareness function, saying at each world which objects the agent is aware of.

In fact this model is in some sense the 'obvious' first-order version of the LGA, at least in the case where full sentential awareness is generated by awareness of atomic formulae. (The main interesting new feature is the possibility of quantification: by using models with non-constant domains, Board and Chung can give a formula corresponding to "The agent is uncertain whether there exists some object she is unaware of", which is a very desirable feature. Even with constant domains, an agent may be uncertain whether any object has the property P while not being uncertain *of* any (particular) object whether *it* has the property P.)

At the same time, the model is given in the 'language' of economics, in terms of events and operators on events rather than a model and an interpreted formal language. This allows a direct comparison with the HMS model, as in a working paper coauthored by originators of both systems [BCS09]. This paper shows, somewhat surprisingly, that OBU structures and HMS structures are to some extent equivalent. That is, if we take a particular generalisation of HMS structures and concern ourselves only with the event structure of the models (ignoring the extra expressive power that quantification brings to the *language* of object-based unawareness), a model in either system can be transformed into a model in the other which captures the same facts about knowledge and awareness of events.

This is surprising because the OBU model, like its close cousin the LGA, allows non-trivial implicit beliefs, while the HMS model does not. The intuitive reason for this surprising result is that the equivalence does not take implicit beliefs into account, as these are nowhere defined in the HMS system.[22] Similarly,

[22] There is a more technical reason also: the HMS models do not entirely follow the description I gave

[HR08] showed that HMS structures can represent exactly the same facts about *explicit* belief and unawareness that the LGA can, but could say nothing about implicit belief.

4.3.1 · *Object-based assumptions*

Since the OBU model is such a close cousin to the LGA, it has roughly the same potential to represent assumptions: we have implicit beliefs, but we cannot distinguish between worlds the agent 'has in mind' but has ruled out and those that she has not even considered. Similarly, there is no way within the theory to relate unawareness to implicit belief: we cannot say which implicit beliefs an agent should hold if she is unaware of some particular object.

As I argued in the first half of this chapter, there *is* no general strategy for deriving implicit beliefs or assumptions from unawareness (remember the assumptions of my key unlocking my own office door but not that of my colleagues). However the particular case of object-based unawareness seems to be different: if I am unaware of some particular *object*, then it seems perfectly reasonable to say that I assume it does not exist. Certainly this works for Walt and the car keys (and possibly even for the interview, if the interview itself is thought of as an object, rather than a proposition "Walt has an interview").

In other words, the object-based model offers the best chance to *derive* assumptions from unawareness. We will still, however, have to do quite a bit of work before we can achieve this: as it stands, the model cannot yet represent the distinction between implicit belief and assumption, which will certainly be needed.

5 · *The rest of the dissertation*

The work on formalising unawareness so far, mainly in the economics community, has concentrated on the distinction between implicit and explicit belief.

above. In the construction of [HMS08], each subspace, corresponding to a vocabulary, contains exactly one world for each maximal consistent set of sentences in that vocabulary. This means that projection downward from the highest subspace forms cones: any set of worlds projects onto just a single world in the lowest subspace with an empty propositional vocabulary. In the construction of [BCS09], on the other hand, each subspace contains *the same number of worlds* as the highest subspace does; downward projection forms *cylinders*, not cones. (I am grateful to Oliver Board for resolving my confusion on this matter.) Worlds in a lower subspace may be identical according to the vocabulary of that subspace, and yet project up to distinct worlds in the highest space. What distinguishes such a pair of worlds cannot be expressed in the logical language of these structures, but from our perspective it is simply *implicit belief*. Of course, in forming the structures this way the connection between the logical language and the semantic representation is broken: two events might be semantically distinct (in terms of the worlds they contain) without there being any formula of the language that can distinguish between them. The careful separation of semantics from syntax may here turn into a drawback rather than an advantage — for instance the canonical model construction of [HMS08], based as it is on the standard technique of identifying states with maximal consistent sets of sentences, cannot produce a state space allowing 'cylindrical projection' in this way.

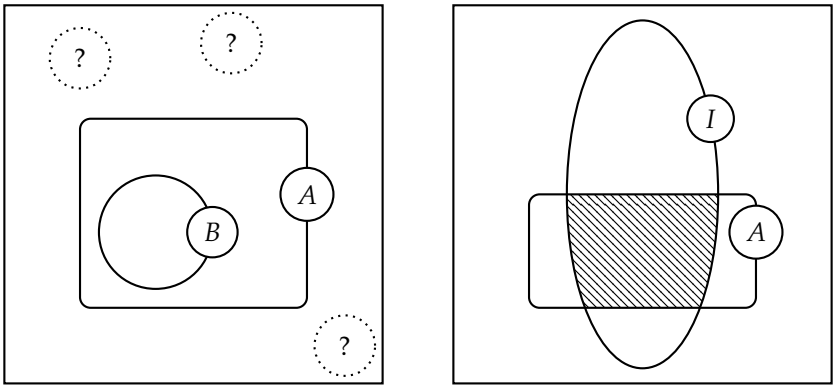
I do not think that this is all there is to the interaction between awareness and belief. An agent's unawareness of possibilities influences her beliefs in systematic ways that go beyond the question of whether she is conscious of them or not.

I have argued, in a sense, for asking the same kinds of questions about an agent's *disbelief*: if a world is not in the agent's belief set, is it explicitly ruled out, or implicitly excluded because the agent does not 'have it in mind'? We need the notion of *assumption*, and we need to ask how assumption interacts with awareness/attention and with belief.

This dissertation is divided into two parts. Figure 1.9 gives an abstract schematic view of the distinction this division embodies. In both pictures the set labelled A represents the worlds the agent 'has in mind'; how this set relates to the agent's (un)awareness of concepts is the fundamental problem of the dissertation as a whole (I make therefore no attempt to answer it in the schematic form of the diagrams).

The distinction between the two parts rests on how we treat the possibilities outside the set the agent 'has in mind'. In Part I (schematically represented in Figure 1.9a), the agent holds definite beliefs about the possibilities within A (the set labelled B) but his attitude to what lies outside remains completely unspecified. If his awareness of possibilities increases, he will have to 'make up his mind' what to believe; the dotted circles represent potential beliefs he might come to hold, but whether these will be realised can only be seen dynamically, as his state of awareness changes over time.

Part II follows a rather different schema, given in Figure 1.9b. As before, A represents the worlds the agent 'has in mind'. He has information which is guaranteed to be objectively true (the set labelled I); however his interpretation of that information (his beliefs, the shaded region) is influenced by his assumptions. In contrast to the models of the first part, from our external perspective we can see how the agent's beliefs will develop as his awareness changes (schematically, his information set I is well-defined even outside the worlds A he 'has in mind'). The interesting feature of these models is rather the relations that are *not* shown in the picture: between his information and his awareness, and his awareness and his assumptions.



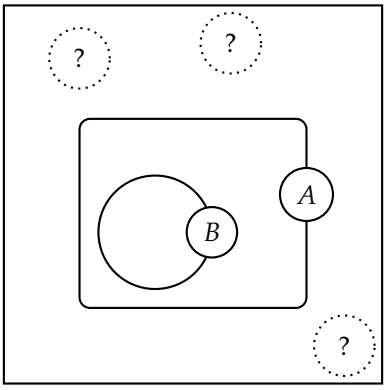
(a) Part I: A is the agent's assumption set, B his belief set. Dotted circles represent possible potential beliefs.

(b) Part II: A is the agent's assumptions, and I his information. The shaded region represents his beliefs.

Figure 1.9: Schematic view of the models of Parts I and II.

Part I

Forming beliefs within assumptions



In these three chapters our agents form assumptions due to unawareness, and beliefs bounded by those assumptions. When their awareness of possibilities changes, so do their assumptions; and they must examine their beliefs anew in light of the new possibilities they come to entertain.

Chapter 2

A model of awareness with assumptions

> [I]gnorance more frequently begets confidence than does knowledge: it is those who know little, and not those who know much, who so positively assert that this or that problem will never be solved by science.
>
> Charles Darwin, *The Descent of Man*

In this chapter I define a simple, static model for representing awareness and assumption. It's a flat (non-relational) single-agent system, which later chapters will equip with dynamics on the model of update semantics [Vel96]. The key idea was introduced in the previous chapter, in my discussion of the logic of general awareness: we need to systematically connect the agent's awareness of atomic formulae (her conceptual vocabulary, or language of self-ascription of belief) to her entertainment of possibilities (the worlds she 'has in mind' when forming her beliefs).

1 · *Some conceptual vocabulary*

Throughout this dissertation I use the tools and notions of possible worlds semantics. A PROPOSITION is a semantic entity, a set of possible worlds. A FORMULA or SENTENCE, on the other hand, is a linguistic entity whose meaning is a proposition. An ATOMIC FORMULA is one that cannot be decomposed into smaller parts (a 'primitive proposition' such as p); I will talk about agents having a CONCEPTUAL VOCABULARY made up of atomic formulae, which can be combined according to the compositional rules of a logical language.

The models I describe have essentially three components, pictured schematically overleaf: the atomic formulae the agent is aware of (her conceptual vocabulary), the worlds she 'has in mind', and the subset of those worlds that she holds possible (her belief set, in the standard sense).

The extra structure provided by the set of worlds she 'has in mind' will come into its own when we define dynamics; in this chapter I want mainly to introduce the technical vocabulary that the rest of the dissertation employs.

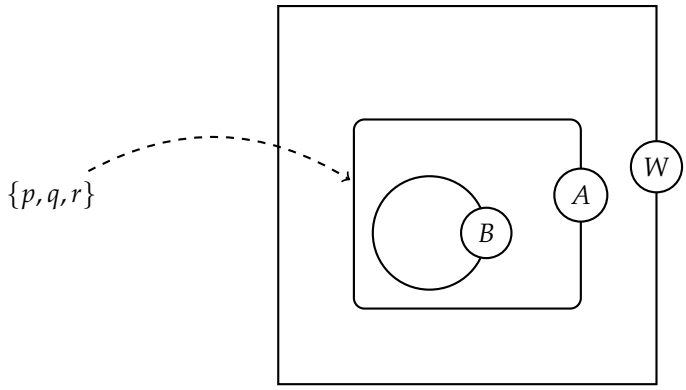


Figure 2.1: Schematic view of the models of this chapter. On the left is the agent's conceptual vocabulary. W is a space of possible worlds, A the worlds the agent 'has in mind', and B her belief set. The models add to the logic of general awareness of [FH88] the set A, and the relationship between that set and the atomic formulae the agent is aware of, indicated by the dashed arrow.

We will distinguish a number of subtly varying propositional attitudes. I am not investigating what we mean when we say "He assumes that φ", rather I have chosen the least inappropriate terms I could find to represent the formal distinctions I want to make. There is some 'slippage', where the common usage of these terms carries connotations that I don't intend, which I try to indicate in this section.

Awareness is not strictly speaking a propositional attitude at all, but an attitude held towards a *formula*. We can be aware of one tautology (I either have or do not have enough money in my wallet for another beer) while not being aware of another (I am either late or not late for class), despite both tautological formulae representing the same proposition. (Note that this means awareness is not closed under substitution of logical identities, nor under entailment.) Awareness is the basis of a distinction that cuts through more or less all instances of the propositional attitudes proper: depending on the particular formula expressing the proposition, the attitude can either be conscious or explicit (the agent is aware of it) or it can be unconscious or implicit (she is not aware of it).[1] As I pointed out in the previous chapter, the term belief is more normally used in ordinary language for *conscious* beliefs;

[1] Following [FH88], in the technical definitions in fact every explicit belief is *also* implicit. As this goes so strongly against normal usage I will tend to highlight the unexpected usage by talking about a (possibly) implicit belief where the inclusive sense needs emphasis.

I will instead use it for both implicit and explicit beliefs. I will also generally elide the distinction between "consciously believes P" (a proposition, a set of worlds) and "consciously believes φ" (a formula, a linguistic expression), wherever it does not seem to risk excessive confusion.

The worlds the agent 'has in mind' are described by the terms ENTERTAINMENT and ASSUMPTION (I call the set variously the 'assumption set' and 'entertainment set', by analogy respectively with 'belief set' and 'possibility set'). She entertains the worlds she 'has in mind', and the propositions that obtain therein; the propositions that hold throughout this set are her assumptions. Most importantly, she can entertain a possibility she does not believe: 'entertainment' is how we can talk about worlds in her assumption set that are not in her belief set.

Assumption and entertainment are modal duals: an agent assumes φ if she does not entertain any non-φ possibilities. Just as it is pragmatically odd to say that someone "holds possible" something they in fact believe, it is odd to say that someone "entertains" a possibility that they in fact assume; nonetheless this is the sense that I mean the terms in, although I will again highlight the pragmatically odd usage when this is important.

Entertainment and assumption, as I use them, are true propositional attitudes: they do not depend on the agent's conceptual vocabulary at all. So (against the more ordinary usage of the term) she need not attend to some formula φ to entertain the proposition it expresses. Typically (although not always) such cases turn on assumption. I do not attend to the continued existence and solidity of my chair when I sit down in it in the morning, for instance: I *assume* that it is solid, and so in all the worlds I 'have in mind' it is solid. Since everything I assume I also entertain (by the pragmatically odd usage), I thus entertain the possibility that my chair is a solid object without attending to it.

The dual of entertainment is ASSUMPTION: I assume φ if I do not represent any possibilities in which φ does not hold. This is clearly what is meant by "I assumed the shop would be open, but I had forgotten it was a bank holiday." However there is another ordinary usage that is *not* intended: "I don't know whether they will be open tomorrow, but I assume they will." The speaker here is referring to what I call simply a belief (in this case an explicit one).

It is *almost* true that assumptions (in the sense I mean) should only ever be unconscious (the agent only assumes (the proposition) φ if she is unaware of (the formula) φ). This gives the right intuition for the shopper: since she is discussing the alternatives in which the shops are shut, she must entertain these alternatives and therefore does not (in my sense) *assume* that the shops are open. However in fact not quite every assumption must be unconscious: the major exception is tautologies. A tautology is of course true in *every* possible world, so it must be true throughout the agent's assumption set; in other words,

she (very reasonably) assumes any tautology φ you like. But she might also be *aware of* φ (say φ is $p \vee \neg p$, and she considers the question whether p holds). The rule is instead that *contingent* assumptions should be unconscious; this is the major principle linking (syntactic) awareness with (semantic) assumption, and recurs in Definition 2.8 below.

So much for the informal notions. Let us turn now to the formal realisation of these idea.

2 · *Syntax*

Many of the languages we consider will contain modal operators. I use the following standard conventions:

NOTATION 2.1: Modal operators. *Modal operators come in pairs of inter-definable duals: if $\Box_X$ is a universal modal with an explicit definition then $\Diamond_X\varphi$ is defined as $\neg\Box_X\neg\varphi$, while if $\Diamond_X\varphi$ is an existential modal with an explicit definition then $\Box_X\varphi$ is defined as $\neg\Diamond_X\neg\varphi$. Typically the set these quantify over will be named X, and X may also stand in as a shorthand operator in the logical language ($B\varphi$ represents belief, for example, and is shorthand for $\Box_B\varphi$, a universal modal interpreted on the agent's belief set, also named B; see below).*

During the course of this dissertation I will introduce a large number of different languages, many of them syntactically related. The following notions, while vague, will help keep track of the variations.

NOTATION 2.2: Object and meta language. *An* OBJECT LANGUAGE *is the kind of language our agents use to communicate with each other; formulae in an object language are the objects of sayings or propositional attitudes. A* META LANGUAGE *is the kind of language we, the modellers, use to describe our agents' thought processes and communications (and possibly actual states of affairs).*

Typically a particular meta language will be defined relative to a particular object language; the meta language of propositional beliefs, for example, has propositional formulae as its object language. But we can also talk about the "meta language of belief", without specifying a particular object language; then we mean something like "The language with a belief operator, under which appear formulae from whichever object language we are interested in at the moment." I will use superscripts to identify object languages and subscripts to identify meta languages. For instance, $\mathscr{L}_B^\Omega$ represents the language of propositional belief: $\mathscr{L}_B$ is the meta language of belief, and $\mathscr{L}^\Omega$ is the propositional object language with non-logical vocabulary Ω.

DEFINITION 2.3: Two object languages: $\mathscr{L}^\Omega$ and $\mathscr{L}^{\Omega,\Diamond}$. *Let Ω be a finite set of atomic formulae; we use these as propositional constants (non-logical vocabulary) to define all our object languages.*

- *$\mathscr{L}^{\Omega}$ is the propositional language with binary conjunction $\wedge$ and unary negation $\neg$, and standard definitions for disjunction, (material) implication, and the biconditional.*
- *$\mathscr{L}^{\Omega,\Diamond}$ adds to $\mathscr{L}^{\Omega}$ a unary prefix operator "might".*[2]

For these and other object languages I will leave off the superscripts (especially Ω) wherever this is unlikely to lead to confusion.

Now we have some examples of object languages, for the use of our agents in communication. Let us see some meta languages, for our use in describing the attitudes of the agents. The following definitions refer to "an" (rather than "the") object language because they can be applied just as they stand to various languages: purely propositional, with "might" or other modalities, with a counterfactual conditional connective, and so on.

DEFINITION 2.4: Two meta languages: $\mathscr{L}_B$ and $\mathscr{L}_A$. *Take an object language $\mathscr{L}$. The* META LANGUAGE OF BELIEF, *$\mathscr{L}_B$, has syntax given by:*

$$\mathscr{L}_B : \xi ::= \Box_B \varphi \mid \xi \wedge \zeta \mid \neg\xi$$
$$\text{where } \varphi \in \mathscr{L} \text{ and } \xi, \zeta \in \mathscr{L}_B.$$

(That is, it is a propositional language whose atoms are belief statements about formulae from the object language.) The less familiar META LANGUAGE OF AWARENESS AND ASSUMPTION, *$\mathscr{L}_A$, has syntax given by:*

$$\mathscr{L}_A : \xi ::= \Box\varphi \mid \Box_A\varphi \mid \Box_B\varphi \mid N\varphi \mid \xi \wedge \zeta \mid \neg\xi$$
$$\text{where } \varphi \in \mathscr{L} \text{ and } \xi, \zeta \in \mathscr{L}_A.$$

For both languages we define $\vee, \rightarrow, \leftarrow, \leftrightarrow$ in terms of $\wedge$ and $\neg$ as usual, and diamond operators as duals of boxes (so $\Diamond_A$ abbreviates $\neg\Box_A\neg$, and so on). In addition, we will use a number of operators more mnemonic than the boxes and diamonds; here is the full list with the intended interpretations:

$A\varphi$ The agent ASSUMES *that φ; shorthand for $\Box_A\varphi$;*[3]

[2] Some logicians might prefer a definition giving a syntactic restriction on combinations of operators, for instance that "*might*" appear only non-nested. Because object languages are intended to be reused, possibly under widely variant semantics, I prefer to leave the syntax relatively unconstrained. I will note when a particular semantics implies some further restriction in order to make sense.

[3] It is unfortunate that all three of "aware", "attend" and "assume" begin with the letter "A". I could not resist the mnemonic that makes A a universal (box) modal and E the corresponding existential (diamond) modal. The reader unfamiliar with the economics literature should bear in mind that A typically means "is aware of", which corresponds to my N (models in the economics literature lack the notion of assumption).

$E\varphi$ *The agent* ENTERTAINS *(the possibility that)* φ*; shorthand for* $\Diamond_A\varphi$*, i.e.,* $\neg\Box_A\neg\varphi$*;*

$N\varphi$ *The agent* ATTENDS *to* φ*.*

$B\varphi$ *The agent* BELIEVES *(possibly implicitly) that* φ*; shorthand for* $\Box_B\varphi$*;*

$X\varphi$ *The agent* EXPLICITLY BELIEVES *that* φ*; shorthand for* $\Box_B\varphi \wedge N\varphi$*;*

This definition places a heavy restriction on the language: no well-formed formula nests operators from $\mathscr{L}_A$, since each must be applied only to a formula from the base propositional language $\mathscr{L}$. This restriction matches the flat semantics, which would give a very unintuitive interpretation to such formulae. (It would of course be an easy matter to extend the syntax, should it become necessary.)

3 · *Semantics*

In a Kripke model for an epistemic modal logic we represent an agent's knowledge by an accessibility relation over worlds; this allows us to represent nested operators (knowledge about knowledge, in a multiagent setting) in a uniform way. Representing the results of epistemic *change* in such a setting is rather complicated, however, and it is in cases of epistemic change that assumptions become most interesting. An alternative representation (based on the Stalnakerian picture of [Sta84] and [Sta99]) models an agent's epistemic state simply as a set of worlds and defines an update function on such states. This representation severely restricts what we can talk about, but its simplicity exposes the basic dynamic structure more clearly.

In a similar manner, I am going to represent a state of attention simply by some concentric sets of worlds (those metaphysically possible containing those entertained containing those believed to be possible). The simplicity comes at a cost: it will not make sense to nest belief *or* entertainment operators. (Although we can represent the fact of assumption or entertainment, we cannot represent beliefs *about* such facts.) However the simplicity of these structures will clearly expose the formal properties of the notions we have just introduced.

The first distinction that formalising these notions makes clear is the need for separate representations of how things 'really are' and of how they appear to the agent. The former contains possible worlds and complete valuations ('what there is'); the latter is a restricted view of what there is, through the agent's 'window of attention'. We will start with the model of what there 'really is'.

DEFINITION 2.5: Metaphysical reality. *Let* W *be a set of possible worlds and* Ω *a finite set of atomic formulae. A* MODEL OF REALITY[4] *is a structure* $M = \langle W, \Omega, V\rangle$ *where* V

[4]This term should not be taken too seriously. Another advantage of the relational semantics is that

associates with each world $w \in W$ a propositional valuation function $V_w \colon \Omega \to \{0,1\}$. We call W the UNIVERSE *of M, and Ω its* VOCABULARY.

The atomic formulae are part of the definition of 'reality' because they are intended to collectively represent all that can be said about the world. In informal discussion I will usually assume that every atomic formula is contingent and that V distinguishes every pair of worlds in W.

Next to the actual state of the world we have the cognitive state of the agent, who may not entertain all possibilities or attend to all possible distinctions between possibilities, and who holds beliefs concerning the worlds she does entertain.

DEFINITION 2.6: Doxastic state. *Let $M = \langle W, \Omega, V \rangle$ be a model of reality. A* DOXASTIC STATE *for some agent is a structure $\sigma = \langle A, B, \Xi \rangle$ where*

$A \subseteq W$ *represents her* ASSUMPTIONS*;*

$B \subseteq A$ *represents her* BELIEFS*; and*

$\Xi \subseteq \Omega$ *generates her language of self-ascription of beliefs, $\mathscr{L}^{\Xi}$.*

This model incorporates one particularly significant simplification: attention to a formula is taken to be at root attention to the propositional constants occurring in that formula. The agent's language of self-ascription of beliefs, $\mathscr{L}^{\Xi}$, is simply the propositional language generated by the atomic formulae the agent attends to.[5]

A model, in this setting, is a pair: metaphysical reality plus a doxastic state. I will reuse the symbol "$\models$" for several truth/support relations: truth at a world, support by a set of worlds, and support by a full doxastic state. Each applies to a different language: purely propositional formulae can be evaluated at single worlds, modal formulae are evaluated on sets of worlds, and formulae from the language of awareness need full doxastic states. The logic is three-valued: $\models$ represents "supports the truth of", $=\!|$ represents "supports the falsity of", and it may be that a given structure supports neither the truth nor the falsity of a given formula. (For example, the doxastic state of an agent uncertain whether p holds supports the falsity of Bp, but the set of worlds in her belief set supports neither the truth nor the falsity of p.)

---

it makes clearer what 'metaphysical reality' in these models really represents: it is simply the state of attention of another agent, namely the agent constructing this particular model (the author).

[5] This is the case [FH88] describes as "propositionally generated partitional awareness". Fagin and Halpern discuss a range of logics similar to this one, in which the agent may attend to any subset of the full language the modeller has at his disposal. The case we use, in which the formulae attended to are generated by a set of atomic formulae, corresponds to the following axioms: $N\varphi \leftrightarrow N\neg\varphi$, $N(\varphi \wedge \psi) \leftrightarrow N\varphi \wedge N\psi$. Since we cannot nest operators we do not represent formulae such as $NE\varphi$.

DEFINITION 2.7: Truth and support relations. *Fix $M = \langle W, \Omega, V \rangle$ a model of reality.*

TRUTH AT A WORLD: *Let $w \in W$ be a world. We evaluate the truth of a formula in $\mathscr{L}^{\Omega}$ as follows:*

$M, w \models p$ $\quad iff_d \quad V_w(p) = 1$
$M, w \models\!\!\!| \; p$ $\quad iff_d \quad V_w(p) = 0$

Standard clauses for $\wedge$ and $\neg$:

$M, w \models \varphi \wedge \psi$ $\quad iff_d \quad M, w \models \varphi$ *and* $M, w \models \psi$
$M, w \models\!\!\!| \; \varphi \wedge \psi$ $\quad iff_d \quad M, w \models\!\!\!| \; \varphi$ *or* $M, w \models\!\!\!| \; \psi$
$M, w \models \neg\varphi$ $\quad iff_d \quad M, w \models\!\!\!| \; \varphi$
$M, w \models\!\!\!| \; \neg\varphi$ $\quad iff_d \quad M, w \models \varphi$

SUPPORT BY A SET OF WORLDS: *Let $S \subseteq W$ be a set of worlds. We evaluate the truth of formulae in $\mathscr{L}^{\Omega,\Diamond}$ as follows:*[6]

$M, S \models p$ $\quad iff_d \quad \forall w \in S : M, w \models p$
$M, S \models\!\!\!| \; p$ $\quad iff_d \quad \forall w \in S : M, w \models\!\!\!| \; p$
$M, S \models$ *might* φ $\quad iff_d \quad \exists w \in S : M, w \models \varphi$
$M, S \models\!\!\!| \;$ *might* φ $\quad iff_d \quad \forall w \in S : M, w \models\!\!\!| \; \varphi$

(and standard clauses for $\wedge$ and $\neg$)

(This definition rules out nesting "might", but allows it to interact freely with other operators.)

METAPHYSICAL POSSIBILITY: *For $\varphi \in \mathscr{L}^{\Omega}$ a (purely propositional) formula:*

$M \models \Diamond\varphi$ $\quad iff_d \quad M, W \models$ *might* φ
$M \models\!\!\!| \; \Diamond\varphi$ $\quad iff_d \quad M, W \models\!\!\!| \;$ *might* φ

SUPPORT BY A FULL DOXASTIC STATE: *Let $\sigma = \langle A, B, \Xi \rangle$ be a doxastic state for M. We evaluate formulae from the (meta) language of awareness $\mathscr{L}_A$ as follows:*

$M, \sigma \models \Box_B\varphi$ $\quad iff_d \quad M, B \models \varphi$ — *belief*
$M, \sigma \models\!\!\!| \; \Box_B\varphi$ $\quad iff_d \quad M, B \not\models \varphi$
$M, \sigma \models \Box_A\varphi$ $\quad iff_d \quad M, A \models \varphi$ — *assumption*
$M, \sigma \models\!\!\!| \; \Box_A\varphi$ $\quad iff_d \quad M, A \not\models \varphi$
$M, \sigma \models N\varphi$ $\quad iff_d \quad \varphi \in \mathscr{L}^{\Xi,\Diamond}$ — *attention to (the formula) φ*
$M, \sigma \models\!\!\!| \; N\varphi$ $\quad iff_d \quad \varphi \notin \mathscr{L}^{\Xi,\Diamond}$

(and standard clauses for $\wedge$ and $\neg$)

In order to keep the definitions as short as possible I have given them for the smallest possible subset of the language. We will mostly use the shorthand notation, though, and it may not be immediately clear how this relates to the formalities above. Here are the more intuitive formulations:

For a purely propositional formula φ, $B\varphi$ is the universal modal on the

[6] This is a static approximation of the dynamic semantics of [Vel96]; the dynamic version will be given in Chapter 3.

B set (belief, possibly implicit) while $A\varphi$ is the universal modal on the A set (assumption); $E\varphi$ (entertainment) is the existential operator on the A set. Modal formulae are simply evaluated on the respective set: a state supports $B(\mathit{might}\ \varphi)$ if its B set supports *might* φ, and so on. (We should not allow modals under E; "entertaining the possibility that φ" should be represented not as $E(\mathit{might}\ \varphi)$ but simply as $E\varphi$, and "entertaining the possibility that $\neg$ *might* $\neg\varphi$" is hard to give a sensible meaning to under the flat semantics.)

The clause for attention is less familiar. It says that the set Ξ of atomic formulae that the agent attends to generates her language of attention $\mathscr{L}^{\Xi,\Diamond}$. If φ is a formula from $\mathscr{L}^{\Omega,\Diamond}$ (that is, possibly containing the modal *might* but without operators such as E or A), then the agent's state supports $N\varphi$ just if φ only uses atomic formulae from the Ξ set of her cognitive state.

In order for these definitions to align with our intuitions not much has to be done (the structure of the models already ensures, for example, that belief is an S4 notion), however we do have to specify some constraints on the relationship between entertainment and attention. A first, alas too simplistic, intuition is that the agent attends to φ if and only if she entertains both φ and $\neg\varphi$. This of course would rule out attention to tautologies, which cannot be right.

Definition 2.8: Attention-consistency. *Let $M = \langle W, \Omega, V\rangle$ be a model of reality and $\sigma = \langle A, B, \Xi\rangle$ a doxastic state. The model M, σ is* attention-consistent *if*

$$\textit{For all } \varphi \in \mathscr{L}^{\Xi}\textit{: if } M \models \Diamond\varphi \textit{ then } M, \sigma \models E\varphi. \tag{2.1}$$

It is belief-attention-consistent *if in addition*

$$\textit{For all } w, v \in A\textit{: if } V_w\restriction\Xi = V_v\restriction\Xi \textit{ then } w \in B \textit{ iff } v \in B. \tag{2.2}$$

The first condition is based on the intuition that attending to φ should involve entertaining both φ and its negation. It says that if φ is metaphysically possible (satisfiable somewhere in M) and the agent attends to φ, then she entertains φ as a possibility. (The condition applies only to propositional formulae, since although the agent can attend to modal formulae these are not satisfied at single worlds but on *sets* of worlds.) This condition corresponds rather obviously to an axiom schema: for all propositional formulae $\varphi \in \mathscr{L}^{\Omega}$, $\Diamond\varphi \rightarrow (N\varphi \rightarrow E\varphi)$.

The second condition ensures that the agent can describe all the (non-assumptive) beliefs she holds. The restriction $V_w\restriction\Xi$ is a propositional valuation for just those atomic formulae in the agent's language of self-ascription of belief. If this language does not distinguish between two worlds w and v, both of which she entertains, then she must not hold different conscious attitudes to them. (What is ruled out is that for some entertained worlds $w, v \in A$ that make the same formulae of L^{Ξ} true, $w \in B$ and $v \notin B$. The agent may *assume* that w and not v is possible, but this is an unconscious attitude.)

Using this notion we can define two entailment relations between sentences: general entailment is the standard notion, but we will mostly be concerned with entailment on the class of attention-consistent models.

DEFINITION 2.9: Entailment. *Let Ω be a set of atomic formulae. Let Γ be a set of sentences and φ a sentence, both from the language $\mathscr{L}_A^\Omega$. We define two notions of entailment:*

$\Gamma \models \varphi$ *iff$_d$ for all models M, σ such that M has vocabulary $\Omega' \supseteq \Omega$: if $M, \sigma \models \gamma$ for each $\gamma \in \Gamma$ then $M, \sigma \models \varphi$;* *(standard entailment)*

$\Gamma \models_A \varphi$ *iff$_d$ for all* attention-consistent *models M, σ such that M has vocabulary $\Omega' \supseteq \Omega$: if $M, \sigma \models \gamma$ for each $\gamma \in \Gamma$ then $M, \sigma \models \varphi$.* *(attention-consistent entailment)*

As usual we write $\models \varphi$ in case Γ is empty.

4 · *Example*

Let us apply this system to describe the various states of mind Walt passes through on his difficult Saturday morning (example 1.1 of Chapter 1).

Let the atomic formulae i, h represent respectively "Walt has a job interview" and "Walt has a hangover". The idea is that Walt is painfully aware of his hangover but (as our story opens) unaware of the interview. We model reality with $M = \langle W, \Omega, V \rangle$ where $\Omega = \{i, h\}$, $W = \{w_0, w_1, w_2, w_3\}$ and V is given by

	i	h
w_0	0	0
w_1	0	1
w_2	1	0
w_3	1	1

Walt passes through two doxastic states: σ_0 represents his state before he attends to the possibility of the interview, while σ_1 represents his state after. When unaware of i he assumes that he does not have an interview, when aware of it he realises that he does. That gives us the values for his belief and assumption sets, and filling in the valuations of the various attitudes is simply a question of definition wrangling.

	A	B	Ξ	Ni	$A\neg i$	$B\neg i$
σ_0	$\{w_0, w_1\}$	$\{w_1\}$	$\{h\}$	0	1	1
σ_1	$\{w_0, w_1, w_2, w_3\}$	$\{w_3\}$	$\{h, i\}$	1	0	0

In the state σ_0, Walt does not attend to the possibility of having an interview; he assumes he has no interview and thus believes it also. After having his attention directed to the interview, he no longer assumes he has no interview;

his belief changes completely, to the belief that indeed he *does*. Throughout this process he continues to realise that he has a hangover.

Both this model and the logic of general awareness of [FH88] represent Walt's belief that he has no interview as *implicit*: he does not attend to the proposition letter i, so he holds no explicit beliefs that mention it. However this model also explains *why* he holds that particular implicit belief: it comes from an *assumption* (the set of worlds he 'has in mind' contains no worlds where he has an interview).

We can compare this assumption (which may be overturned) to his belief that he has a hangover (which may not be overturned), by looking at his attitude to the two worlds w_0 and w_3, when he is in the state σ_0. Neither world features in his belief set, but for two quite different reasons: w_0 is excluded because he has consciously considered it and believes it is not actual (in that world he had no hangover), while w_3 is excluded only by his assumption. This difference in turn shows itself in the behaviour of the two worlds under the update: w_3 *may* enter his belief set (since it was hitherto excluded by assumption) while w_0 may *not* (since it has been considered and rejected).

5 · *Properties*

The most interesting properties of this system will emerge in the following chapter, when we introduce the update formalism for attention dynamics. A few properties can already be described, however.

For instance, $A\varphi \models B\varphi$: an agent cannot disbelieve her assumptions. (This holds for all models, since $B \subseteq A$.) Equally, though, $A\varphi \wedge \Diamond\neg\varphi \models_A \neg X\varphi$: if the agent holds a contingent assumption, her belief in that assumption is strictly implicit (this follows from condition (2.1) on attention-consistent models; note that attention-consistent entailment is used).

Another direct consequence of this condition is that $N\varphi \models_A \Diamond\varphi \leftrightarrow E\varphi$: if φ is in the agent's language, then she entertains it iff it is metaphysically possible. That is, entertainment captures the *agent's* view of what is metaphysically possible.

We cannot directly compare the expressivity of this system with that of the LGA (the logic of general awareness of [FH88], discussed in Section 4.1 of Chapter 1): the LGA allows nested belief operators while this language includes the assumption operator which has no counterpart in the LGA. However we can get an intuitive picture of the relationship by looking at the 'common core' of the two systems. Suppose we take only the operators N, B, and X (restricting to the language of the LGA), consider only formulae without nested operators (restricting to the language of flat models), and compare a single doxastic state (as defined in this chapter) with a single-agent pointed LGA model in which all worlds accessible from the actual world w see each other and the agent is aware of the same formulae at all worlds (the equivalent within a relational system of

the 'information state' flat semantics: the connected component visible from w corresponds to the belief set B). Now we can ask, does every such restricted model for the LGA have a corresponding model and doxastic state (in the sense of this chapter) making the same restricted set of sentences true, and vice versa?

The answer in both cases is yes. It is easy to construct the restricted LGA model corresponding to a doxastic state: simply throw away the assumption set A and make all worlds in the belief set see each other.

Going back the other way is more complicated. The accessible worlds must of course correspond to the belief set B, but the question of what to do with the assumption set A still remains. Not every choice will be acceptable: the LGA model fixes the B-worlds and the agent's language of attention, so not every choice of A will produce a belief-attention-consistent state.

Here is an example. Suppose our models contain only two worlds:

	p	q
w_0	1	1
w_1	1	0

Take two models of the LGA, in both of which the belief set of the agent is just $\{w_0\}$. In M_1 she is aware of q but not p, in M_2 she is aware of p but not q.

The definition of belief-attention consistency constrains us to a particular decision about the assumption set A in each of these models: in M_1 it must be $\{w_0, w_1\}$, while in M_2 it must be $\{w_0\}$.

In M_1 the agent is aware of q; attention-consistency requires that she entertain all valuations of q that occur in the model. Both possible valuations do occur, so both w_1 and w_2 must be included in her assumption set.

In M_2, on the other hand, the decision is forced by *belief*-attention consistency. If w_1 were included in A, then the agent's non-assumptive beliefs would distinguish between w_0 and w_1; but she is aware only of p, so she cannot explain what it is about w_1 that distinguishes it from w_0 and justifies her belief.

We might wonder if it is always possible to find an assumption set satisfying the requirements of belief-attention consistency. In fact it is. The following construction ensures it: A world belongs in the set A if and only if either

1. it is in B, or
2. it differs from every world in B on the valuation of at least one atomic formula in Ξ.[7]

[The second condition directly establishes the extra condition for belief-attention consistency, so we need only check that the resulting state is attention-consistent. If it is not, then there is some valuation of the atomic formulae in Ξ

[7] B is nonempty because the accessibility relation for the LGA has no dead ends; the special case with empty B is anyway easily accommodated by taking A to be the full set W of worlds.

that is witnessed at a world w in the model, but not witnessed anywhere in A. Since $B \subseteq A$, that means no worlds in B have this valuation of the formulae in Ξ. But then w differs from every world in B on its valuation of some atomic formula in Ξ, so the second condition ensures that $w \in A$.]

This means that in one sense the model adds nothing to the LGA except for the notion of assumption. The assumptions that are added, though, are not free to vary entirely as they like: they must respect the constraints that relate them to the agent's beliefs and awareness. In the next chapter we will see how these constraints operate across dynamic awareness updates. The chief function of the assumption set (besides its value as an intuitive notion) is to allow us to express the right conditions on updates which may change the agent's beliefs or awareness of possibilities or both.

Chapter 3

Attention dynamics

> And as soon as she had asked herself the question, she knew the answer.
>
> Neil Gaiman, *Coraline*

> It's the difficult thing, the hard thing, about secrets. Sometimes you really don't want to know them. But once you do, there's no going back out; no unlearning them.
>
> Aunt Cloud in John Crowley's *Little, Big*

The flat semantics of the previous chapter makes a distinction that the logic of general awareness of [FH88] cannot: between worlds that are *consciously ruled out* and those that are *assumed away*. This distinction might be relatively unimportant when we look only at a static situation: what matters for explaining the behaviour of the agent is simply the worlds in her belief set, which is equally simply defined in either system.[1] When we start thinking about how an agent might *update* her cognitive state, though, this distinction becomes essential. It is to this problem that we now turn.

Frank Veltman's [Vel96] sets the standard for update languages. In the following section I will describe his semantics in some detail, although for full motivation I refer the reader to the original paper.[2] My own approach will incorporate Veltman's semantics without change, but embedded in a larger structure: his semantics for *belief* update will be augmented with a representation of *attention* update.

1 · *Update languages*

Definition 3.1: Update system. *An* update system *is a triple* $\langle \mathscr{L}, \Sigma, [\cdot] \rangle$ *where*

- $\mathscr{L}$ *is a language (in my terminology an object language);*
- Σ *is the set of possible information states; and*
- $[\cdot]$ *is a function taking each formula to an* update function *on states: if* $\varphi \in \mathscr{L}$ *is a formula then* $[\varphi] : \Sigma \to \Sigma$ *is the* update function *for* φ.

[1] Indeed from such a perspective the representation of nested (multi-agent) belief, sacrificed in our flat semantics, will probably appear much more important than this subtle distinction.

[2] I discuss only the first system described in [Vel96], with semantics for *might*. I make no use of the default reasoning introduced in the latter part of the paper; the approach is I think entirely consistent with, but orthogonal to, my account.

Update functions are written in postfix notation; if σ is an information state then $\sigma[\varphi]$ is the information state resulting from updating σ with φ (the function $[\varphi]$ applied to σ). This is convenient when we want to represent sequences *of updates: $\sigma[\varphi][\psi]$ is read as "σ updated with φ, then the result updated with ψ".*

The update system we are interested in is defined on the object language $\mathscr{L}^{\Omega,\Diamond}$ (that is, a propositional language with modal operator *might*). This is the dynamic system that the static semantics for *might* of the previous chapter is based on.

DEFINITION 3.2: Veltman's update system for *might*. *Fix a model of reality, $M = \langle W, \Omega, V\rangle$. Veltman's update system is $\langle \mathscr{L}^{\Omega,\Diamond}, \mathscr{P}(W), [\,\cdot\,]^M_{\mathrm{V}}\rangle$, with the components defined as follows:*[3]

The language $\mathscr{L}^{\Omega,\Diamond}$ is the propositional language with "might".[4]

An information state is a set of possible worlds: an element of $\mathscr{P}(W)$.

The update function for $[\varphi]^M_{\mathrm{V}}$ is defined by recursion on the structure of φ, as follows:

$$
\begin{aligned}
\sigma[p]^M_{\mathrm{V}} &=_d \sigma \cap \{w \in W \,;\, V_w(p) = 1\} \\
\sigma[\neg\varphi]^M_{\mathrm{V}} &=_d \sigma \setminus \sigma[\varphi]^M_{\mathrm{V}} \\
\sigma[\varphi \wedge \psi]^M_{\mathrm{V}} &=_d \sigma[\varphi]^M_{\mathrm{V}} \cap \sigma[\psi]^M_{\mathrm{V}} \\
\sigma[\varphi \vee \psi]^M_{\mathrm{V}} &=_d \sigma[\varphi]^M_{\mathrm{V}} \cup \sigma[\psi]^M_{\mathrm{V}} \\
\sigma[\textit{might}\ \varphi]^M_{\mathrm{V}} &=_d \begin{cases} \varnothing & \textit{if } \sigma[\varphi]^M_{\mathrm{V}} = \varnothing \\ \sigma & \textit{otherwise} \end{cases}
\end{aligned}
$$

Precision demands the profusion of superscripts identifying the model, at least in definitions; I will leave them off almost everywhere in the text, as this is highly unlikely to cause confusion.

All the clauses of this definition are classical except for *might*. The update with "*might* φ" is of a special kind, known as a TEST: either it leaves the information state unchanged (the test SUCCEEDS) or it takes it to the empty (absurd) state (the test FAILS). We will call an update that (on some information state) is not a test INFORMATIONAL or SUBSTANTIVE.

DEFINITION 3.3: Acceptance. *A formula φ* IS ACCEPTED IN *a state σ, written $\sigma \Vdash \varphi$, iff $\sigma[\varphi] = \sigma$.*

(This is not the notation Veltman uses. I use $\Vdash$ to distinguish this dynamic notation from the support relation $\models$ of the previous chapter.)

[3] The subscript V stands for "Veltman"; we will be defining several more update functions.

[4] There are minor differences of detail between my formulation and Veltman's, most of which need not concern us here; in particular, Veltman's system can deal with some constructions involving nested operators that mine cannot.

1.1 · *Lifting to cognitive states*

We are going to define an update language based on Veltman's, but interpreted on the more extensive cognitive states introduced in the previous chapter (incorporating both awareness and belief). Such a state has three components: $\langle A, B, \Xi \rangle$, corresponding to the agent's *assumptions*, *beliefs*, and her *language of attention*. It is easy to see the place that Veltman's definition should have in such an account: it corresponds to updates of the *belief component* B of the state.

Definition 3.4: Pure belief update. *Fix a model of reality, $M = \langle W, \Omega, V \rangle$. Let Σ^M be the set of cognitive states for M. The update system of* pure belief update *(for M) is $\langle \mathscr{L}^{\Omega}_{\mathrm{V}}, \Sigma^M, [\,\cdot\,]^M_{\mathrm{b}} \rangle$, where the update function is given by*

$$\langle A, B, \Xi \rangle [\varphi]^M_{\mathrm{b}} =_d \langle A, B[\varphi]^M_{\mathrm{V}}, \Xi \rangle.$$

Simple though this formal lift is, it has far-reaching conceptual consequences. For Veltman, acceptance is a *normative* notion: "if $\sigma[\varphi]_{\mathrm{V}} = \sigma$, an agent in state σ *has* to accept φ" [Vel96, p. 229, orig. emph.]. This is certainly no longer the case according to the definition I have just given, for reasons that (of course) hinge on unawareness. To see this, it is time to turn to examples. Let me introduce a scenario which will be developed (in increasingly ridiculous directions) throughout the rest of the chapter.

Example 3.5: Olga the ornithologist. *On a fine spring day in Amsterdam, Olga the ornithologist is enjoying the afternoon sun when she sees a distant object in the sky. It so happens that the three kinds of bird that she attends to are ducks, herons, and ravens. She can't tell which of these the distant object is, but it is fairly light-coloured; she thinks it might be a duck or a heron but is sure it can't be a raven.*

Figure 3.1a overleaf shows this state; the rest of Figure 3.1 shows the effect of various pure belief updates.

The three figures in the top row presents no problems: hearing "d" she gains information, while hearing "*might* d" she learns nothing new (the test succeeds). All three pictures of the bottom row represent the absurd belief state, but for three different reasons. In 3.1d the attempted update directly contradicts the information Olga already has; to resolve the issue she will have to decide whether her information or that of her informant is more reliable, a process (probably involving some kind of belief revision) which I will not attempt to model. In 3.1e Olga's situation is somewhat easier: she must still reject the update, but instead of needing belief revision to resolve the conflict between her an her informant, she just needs to tell the informant that her own information is better: "It's not a raven, look, it's white."

So far everything has gone according to plan, following Veltman's scheme; however the final figure, 3.1f, reveals a problem. Intuitively the goose possibility ought to be compatible with Olga's state (the test ought to succeed, rather than

fail), however this can only be the case if her attention set A is extended *before* the pure belief update is applied. This is the reason for the emphatic nomenclature of Definition 3.4: $[\varphi]_{\mathrm{b}}$ is a *pure* belief update in the sense that it does not take any account of awareness issues. While the notion is formally very useful, I suspect that pure belief updates hardly ever occur 'in the wild', and certainly never as the result of utterances in conversation. Most beliefs we form (and all those that come from processing some linguistic utterance) are first *conceptualised*, and only then incorporated into the general belief state. The conceptualisation step brings with it an *awareness* update, complementing Veltman's belief update and completing the picture.

2 · *Updates with awareness*

If Veltman's notion of acceptance is no longer normative when lifted into cognitive states with awareness, what is it? In fact, it corresponds exactly to a notion introduced in the previous chapter: σ accepts φ just in case $\sigma \models B\varphi$ according to the static semantics given in Definition 2.7. That is, acceptance corresponds to (possibly *implicit*) belief; it is the potential for implicit belief that provides the interaction with awareness. Given this parallelism, what we would like to do is define an attention update, say $[\,\cdot\,]_{\mathrm{n}}$, such that acceptance according to that update corresponds to $\sigma \models N\varphi$ in the static semantics; then the effect of hearing an utterance of φ would be simply the awareness update followed by the belief update. (The subscript is "n" rather than "a" for awareness because it corresponds to the operator $N\varphi$ rather than $A\varphi$ for assumption.) That is what we will do in the rest of this section, although the awareness update turns out to be surprisingly involved.

Definition 3.6: Update systems with awareness. *Fix a model of reality, $M = \langle W, \Omega, V\rangle$. Let Σ^M be the set of belief-attention-consistent cognitive states for M. We define three interrelated update functions, for belief, awareness, and utterances.*

Let $\sigma = \langle A, B, \Xi\rangle$ be a cognitive state, and φ a formula.

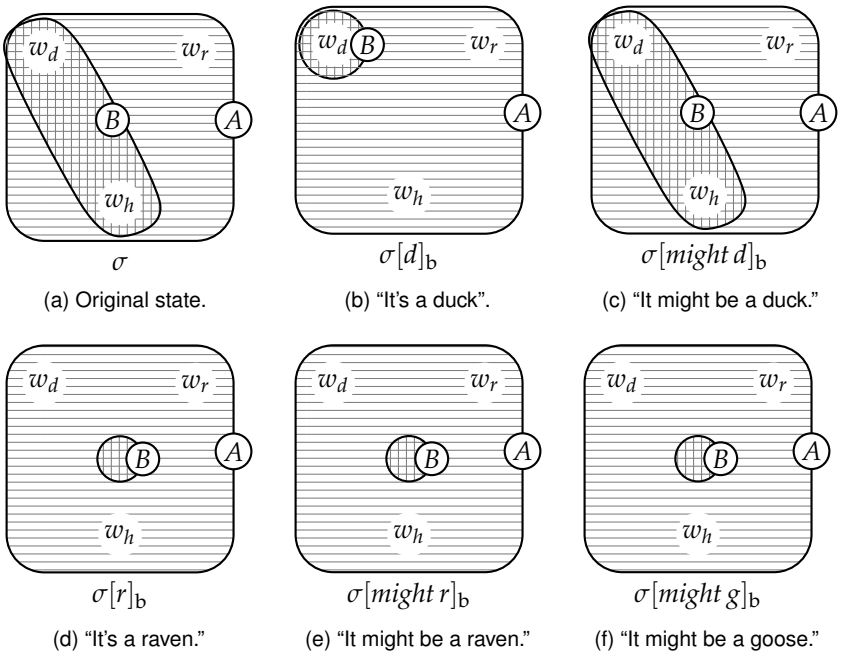
$[\varphi]^M_{\mathrm{n}}$ represents an update with pure awareness *of φ. This is the most complicated of the three, and I will defer the full definition for a moment. Intuitively, though, this corresponds to nothing more than becoming aware of φ; in particular, no new information is gained about the worlds already in the awareness set A. Becoming aware of φ will certainly add all proposition letters occurring in φ to Ξ, however, as well as potentially adding* new *worlds to A.*

$[\varphi]^M_{\mathrm{b}}$ represents an update with pure belief *that φ, as given in Definition 3.4:*

$$\sigma[\varphi]^M_{\mathrm{b}} =_d \langle A, B[\varphi]^M_{\mathrm{V}}, \Xi\rangle.$$

$[\varphi]^M$ represents an utterance update *with φ. The definition is simple:*

$$\sigma[\varphi]^M =_d \sigma[\varphi]^M_{\mathrm{n}}[\varphi]^M_{\mathrm{b}}.$$



(a) Original state. (b) "It's a duck". (c) "It might be a duck."

(d) "It's a raven." (e) "It might be a raven." (f) "It might be a goose."

Figure 3.1: Olga the ornithologist, with pure belief updates. Horizontal shading indicates the A set, vertical shading the B set. The worlds are named mnemonically: in w_d the bird is a duck, in w_r a raven, and in w_h a heron.

Given the conceptual differences between Veltman's system and mine, I will talk about "support" rather than "acceptance". (Note however that the support relations $\models$ and $\Vdash$ are formally distinct; "support" means the latter in this chapter, unless otherwise specified.)

Definition 3.7: Support with awareness. *Fix M as before. Let σ be a cognitive state and φ a formula from $\mathscr{L}^{\Omega}_{\mathrm{V}}$. Then*

$$\begin{aligned}
\sigma &\Vdash^{M} N\varphi \textit{ iff}_{d}\ \sigma[\varphi]^{M}_{\mathrm{n}} = \sigma \\
\sigma &\Vdash^{M} B\varphi \textit{ iff}_{d}\ \sigma[\varphi]^{M}_{\mathrm{b}} = \sigma \\
\sigma &\Vdash^{M} X\varphi \textit{ iff}_{d}\ \sigma[\varphi]^{M} = \sigma
\end{aligned}$$

The three modalities are respectively attention, (possibly implicit) belief, *and* explicit belief. *We can also capture* assumption, *since that is equivalent to belief*

evaluated over the set A:

$$\langle A, B, \Xi \rangle \Vdash^M A\varphi \textit{ iff}_d \langle A, A, \Xi \rangle \Vdash^M B\varphi$$

All that remains to fill out this picture is a definition of the awareness update. However this is less simple that it might at first appear. To start with the easy part: becoming aware of φ adds all the proposition letters occurring in φ to the agent's set Ξ (the atomic formulae she attends to). The complications arrive because both the A set *and* the B set may need to be adjusted in light of the new awareness; the A set because considering new possibilities naturally makes the agent entertain new worlds, and the B set because these new worlds may be plausible enough for her to hold them possible. At the core of the update, however, is the change to Ξ; the definitions I will propose make all other changes dependent on that.

Before giving the definitions themselves, let us see a few examples of pure awareness updates (if only to show that all is not as smooth and simple as might be hoped).

2.1 · *Desiderata for the awareness update*

Figure 3.2 on the facing page gives a number of awareness updates (some more sensible than others) showing the range we have to cover.

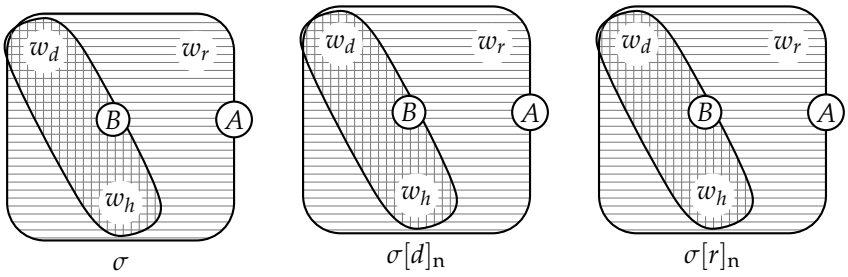
Again, the first row is unproblematic: if Olga already attends to d and r then (regardless of her beliefs about them) the awareness update should be vacuous.

In the second row some new worlds have to be added to Olga's A set; however w_g is treated differently to w_c. Remember that Olga can see that the bird is white; while she will *entertain* both goose- and crow-worlds if they are brought to her attention, she should only hold the goose possible.

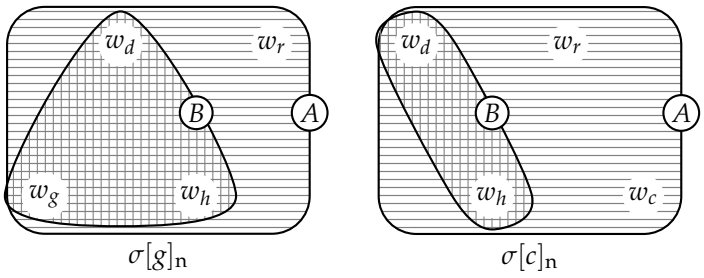
In the third row things get somewhat ridiculous, I must admit; the point of the exercise is to observe that there are still regularities in the effects of awareness updates even when the possibilities they bring to light are ones we do not wish to take seriously. In Figure 3.2f Olga considers the possibility that she is seeing a goose that is also the incarnation of Zeus. Let us assume that our model of reality admits this as a metaphysical possibility (Zeus seduced Leda in the form of a swan, after all). Several important things happen: Olga entertains the world w_g (whereas in 3.2g, where she does not attend to g, she does not); she also entertains a number of worlds where Zeus impersonates the various other birds she attends to; and she forms sensible (i.e., non-Zeus) beliefs about all of these.

Taking all of these updates together, we can establish several desiderata for the update mechanism.

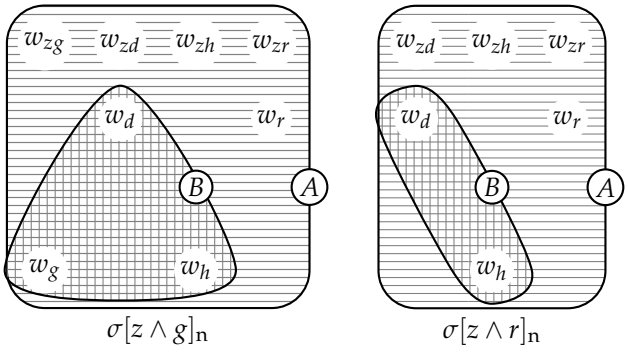
1. Adding worlds to A. Thinking about g brings the world w_g to mind. It



(a) Original state (b) Thinking about a duck (c) Thinking about a raven



(d) Thinking about a goose (e) Thinking about a crow



(f) Thinking about Zeus disguised as a goose (g) Thinking about Zeus disguised as a raven

Figure 3.2: Olga the ornithologist, with pure awareness updates. Horizontal shading indicates the A set, vertical shading the B set. The worlds are again named mnemonically; d for duck, r for raven, h for heron, c for crow, g for goose, and z for Zeus disguised as some kind of bird (for the unexpected intrusion of the Greek deity see the text).

does not, however, add to the A set every world where g holds; w_{zg} is one such, which we would not want added unless z is explicitly under consideration.

2. Plausible new worlds may be held possible. Thinking about g, the new world w_g gets added to the B set.

3. Implausible new worlds need not be held possible. Thinking about c, the world w_c is not added to B; neither are the various Zeus possibilities.

4. Combinatorics. Thinking about Zeus impersonating one bird naturally brings to mind impersonations of others.

To these must be added one more, with a more theoretical basis:

5. The update must preserve belief-attention consistency.

Recall the notions of attention and belief-attention consistency from the previous chapter (Definition 2.8): a state is attention-consistent if the agent entertains a world for each metaphysically possible valuation of the proposition letters she attends to, and it is belief-attention consistent if in addition her language of awareness suffices to describe her beliefs within her A set. We want the utterance update $[\varphi]$ to preserve belief-attention consistency (if σ is belief-attention consistent, so is $\sigma[\varphi]$ for all φ); to ensure this, it is both necessary and sufficient that $[\varphi]_{\mathrm{n}}$ preserve belief-attention consistency.

[Necessity: Let σ be belief-attention consistent and $\sigma[\varphi]_{\mathrm{n}}$ not. Both φ and $\varphi \vee \neg\varphi$ contain the same proposition letters, so induce the same awareness updates; and the belief update with $\varphi \vee \neg\varphi$ is vacuous. Thus:

$$\begin{aligned}\sigma[\varphi \vee \neg\varphi] &= \sigma[\varphi \vee \neg\varphi]_{\mathrm{n}}[\varphi \vee \neg\varphi]_{\mathrm{b}} \\ &= \sigma[\varphi]_{\mathrm{n}}[\varphi \vee \neg\varphi]_{\mathrm{b}} \\ &= \sigma[\varphi]_{\mathrm{n}}.\end{aligned}$$

If the latter is not belief-attention consistent, neither is the former, and $[\,\cdot\,]$ does not preserve belief-attention consistency. For sufficiency, note that $[\varphi]_{\mathrm{b}}$ only separates worlds (in or out of B) that differ on some proposition letters in φ; but it is preceded by an awareness update with φ, adding exactly those letters to Ξ, so the agent has the vocabulary to describe the separation.]

We can separate the desiderata into two mechanisms: both the A set and the B set need to be updated. Each mechanism involves both choice and constraint: we can choose new worlds to add to A, but desiderata 1 and 4 come down to requiring that the update preserve attention consistency; of the new worlds, we can choose which ones end up in B, but our choice must preserve belief-attention consistency.

Let us begin with the easier of the two: the update to A. I will introduce a component sitting conceptually somewhere in between the model of reality and the agent's cognitive state; like the state it is personal to the agent, but like the model it is immutable and constrains the evolution through states produced by updates. I call it the agent's PERSONALITY.

2.2 · *Agents with personality*

We saw that when Olga becomes aware of g she should come to entertain some, but not all, of the worlds where g holds; the first question is, which ones? Looking at the example of Zeus, the answer seems to be something like "the most sensible ones". We will represent this with a simple ordering: if Olga needs some new worlds where some formula φ holds, she will take the *least* worlds satisfying φ in the ordering.

The notion of course comes from Lewisian counterfactual semantics, where a similarity ordering compares how similar w_1 and w_2 are to a reference world w; here we have no reference world, and we compare rather how 'normal' w_1 and w_2 are in terms of the agent's expectations.

So let us say that $\preceq$ is a weak order on W (a reflexive, transitive relation such that for each $w, v \in W$ at least one of $w \preceq v$ and $v \preceq w$ holds; also known as a 'complete preorder'). I will call it the ASSOCIATION ORDERING. We can visualise this as a 'system of spheres' or as a linear order of equivalence classes.

Define the $\preceq$-minimal worlds in W satisfying φ as

$$\min_{\preceq} W\restriction\varphi =_{\mathrm{d}} \{w \in W; M, w \models \varphi \wedge \neg\exists w' \in W\colon M, w' \models \varphi \wedge w' \prec w\}.$$

Then if Olga needs some new worlds satisfying g, she should add

$$\min_{\preceq} W\restriction g$$

to her A set.

The question arises, where should we put this ordering $\preceq$? It is particular to the agent (it represents what worlds spring to *her* mind), so perhaps it should go in her cognitive state. But unlike the other components of the state, it is immutable; the cognitive state evolves over time, while the association ordering does not. We will add it as a separate index: the update of a state is calculated relative to a model *and a personality*.

DEFINITION 3.8: Personality (preliminary). *Fix a model $M = \langle W, \Omega, V\rangle$. An agent's* PERSONALITY *is a structure $\Pi = \langle \preceq, \ldots\rangle$ whose first component is a weak order on W; the other components are given in Definition 3.14, and have to do with the belief-change part of the awareness update.*

Actually we can say more than just how the personality affects an update. If you imagine that the agent's current state is the outcome of a prior series of

updates, it seems that once the personality is fixed, not every cognitive state should be reachable. The following definition captures this notion, in such a way that the A set is *completely derived* from the agent's personality and the proposition letters she attends to.

DEFINITION 3.9: Π-consistent state. *Fix M a model of reality and $\Pi = \langle \preceq, \ldots \rangle$ a personality. A cognitive state $\sigma = \langle A, B, \Xi \rangle$ is* Π-CONSISTENT *if it is belief-attention consistent and either*

(a) $\Xi = \emptyset$ and $B = A = \min_{\preceq} W$; or

(b) $\Xi \neq \emptyset$ and

$$A = \bigcup \left\{ \min_{\preceq} W{\restriction}\varphi \; ; \; \varphi \in \mathscr{L}^{\Xi} \right\}.$$

This definition is convenient in two ways. It lets us 'read off' an agent's A set from her personality and the proposition letters she attends to (thus indirectly fulfilling desideratum 1 above). But it also resolves an embarrassing corner case: what should an agent entertain when she attends to *no* proposition letters at all? The answer Definition 3.9 gives is that the lowest rank of $\preceq$ provides the agent's most basic assumptions, which seems natural. And it is easy to see, by the very construction, that the resulting A set will always be attention-consistent with Ξ.

We are halfway to having defined our update. The next task is to describe the strategy Olga uses to form beliefs about newly noticed possibilities, so that we can distinguish between the goose (possible) and the raven and Zeus (respectively ruled out and ridiculous).

2.3 · *Spontaneous belief formation*

The mechanism we will define is a part of the agent's personality; it is a static specification of her propensity to form spontaneous beliefs under changing conditions.

The first thing to notice about this mechanism is that it must operate only on the worlds newly added to A: it is responsible for deciding which of those *new* worlds go into B and which do not, but not for moving worlds that were already in A into or out of B.

The reason is that the boundary between B and A is where the agent accumulates knowledge taken from the dialogue. The function of the spontaneous belief formation component is to represent the agent's own 'background knowledge' (that crows are black, that the bird she sees is white, that Zeus doesn't manifest these days), and the way this influences her when she considers a new possibility. Its function is not to allow her to reconsider previous decisions in the light of new awareness.[5]

[5]Previously-held beliefs may be overturned, but this still works by the addition of *new* worlds; no

So we can describe the functioning of this mechanism like this: given the set of new worlds added to A by the awareness update, it decides which of them will go into B and which will not. Now how should this decision be defined?

The simplest thought might be just to specify a 'background belief' subset of W: worlds in that set go into B, worlds outside it don't. This won't get us very far, though, because the resulting state needs to be belief-attention consistent. Without complicated constraints relating $\preceq$ to the background belief set, it would be possible for the agent to form a belief she could not justify using her language $\mathscr{L}^{\Xi}$; even if such constraints could be written down, it seems very likely to me that they would trivialise the resulting beliefs (that is, the belief set would have to match the structure of $\preceq$ so closely that it might as well not exist as a distinct structure).

A similar objection applies to the next most obvious possibility: that spontaneous beliefs are generated via an order like $\preceq$, with the minimal worlds of the new set entering B. If the ordering separates worlds that are not separated by $\preceq$ then careful manipulation can produce a state that is not belief-attention consistent; if it does not, then the beliefs formed will always be trivial (in this formulation, all new worlds will enter B; various alternatives exist but I haven't found any that do any better, nor do I expect anyone else to do so).

The problem with both these suggestions is that they do not pay sufficient respect to the influence of Ξ in the formulation of belief-attention consistency. In the interests of simplicity I am going to propose an admittedly deficient solution to this problem: we will generate the spontaneous beliefs with a function that takes the Ξ set as a parameter, as well as the set of new worlds, and which then can be guaranteed to produce only belief-attention consistent beliefs.[6]

Call the function forming spontaneous beliefs $\mathfrak{B}$ (deferring for a moment the constraints that will make it well-behaved). It belongs in the agent's personality (which will also be given its final definition in a moment); with its help, we can now define the pure awareness update.

---

previous uncertainties can be resolved by a pure attention update. Of course this is a simplification in the service of abstraction; in the probabilistic setting I describe in Chapter 5 it no longer holds entirely true.

[6] This solution is unsatisfying in that certainly not just any function will do; I will have almost nothing to say about possible constraints the function should obey. If we were to write down a particular example of such a function, we would quickly notice another representation mismatch: in specifying the function we would have to describe what the agent should believe about sets of worlds that no update will ever deliver to the function, because of the structure of her association order. Finally, there may be examples we would like to represent which this strategy will not cover; it may be, for example, that spontaneous beliefs should vary depending on *current* beliefs, or the worlds currently entertained, or some such. However these are all either methodological objections or entirely speculative. In the absence of intuitive examples showing that this approach is deficient, and in the interests of simplicity, this will do for now.

DEFINITION 3.10: Pure awareness update. *Fix M a model and $\Pi = \langle \preceq, \mathfrak{B} \rangle$ the agent's personality. Let $\sigma = \langle A, B, \Xi \rangle$ be a Π-consistent cognitive state. Then $\sigma[\varphi]_{\mathrm{n}} =_d \langle A', B', X' \rangle$ where*

- $\Xi' = \Xi \cup \{p \in \Omega \,;\, p \text{ occurs in } \varphi\}$ *(the agent comes to attend to all proposition letters occurring in φ);*
- *A' is generated by $\preceq$ and Ξ':*

$$A' = \bigcup \left\{ \min_{\preceq} W\restriction\psi \,;\, \psi \in \mathscr{L}^{\Xi'} \right\};$$

- $B' = B \cup \mathfrak{B}(A' \setminus A, \Xi)$ *(the agent's spontaneous beliefs about the* new *worlds she entertains are given by her selection function).*

Note that the update according to this definition will actually preserve Π-consistency if it preserves belief-attention consistency; this is of course what we want.

So what constraints on $\mathfrak{B}$ are required if the new state is to be belief-attention consistent? A failure of consistency requires two worlds $w, v \in A'$ such that $V_w\restriction\Xi' = V_v\restriction\Xi'$, but (say) $w \in B'$ but $v \notin B'$. Since $\Xi' \supseteq \Xi$ and σ was belief-attention consistent, it cannot be that $w, v \in A$; at least one of the two must be new. The following lemma shows that in fact neither w nor v can have been present in A.

LEMMA 3.11: Imagination. *If $w \in A$ and $v \in A' \setminus A$, then $V_w\restriction\Xi' \neq V_v\restriction\Xi'$. Equivalently, every new world in A' satisfies some formula of $\mathscr{L}^{\Xi'}$ that was not satisfied anywhere in A.*

Proof. Suppose otherwise: v and w satisfy exactly the same formulae in $\mathscr{L}^{\Xi'}$. Since $v \in A'$, for some $\varphi \in \mathscr{L}^{\Xi'}$, v is a minimal world satisfying φ; likewise for w and some $\psi \in \mathscr{L}^{\Xi}$. By hypothesis the worlds satisfy the same formulae of $\mathscr{L}^{\Xi'}$, and $\mathscr{L}^{\Xi'} \supseteq \mathscr{L}^{\Xi}$, so v also satisfies ψ. But v is not a *minimal* world satisfying ψ (or it would have been in A), so $w \prec v$. But since w and v satisfy the same formulae, w satisfies φ and $w \preceq v$; v is therefore not a minimal world satisfying φ. But this is a contradiction, so v and w cannot satisfy exactly the same formulae in $\mathscr{L}^{\Xi'}$. qed

This means that to avoid belief-attention-inconsistent pairs w, v we need only look within $A' \setminus A$. Now we can simply require that $\mathfrak{B}$ is well-behaved in the right way:

DEFINITION 3.12: Awareness-relative selection functions. *Let $M = \langle W, \Omega, V \rangle$ be a model. A function $f \colon \mathscr{P}(W) \times \mathscr{P}(\Omega) \to \mathscr{P}(W)$ is an* AWARENESS-RELATIVE

SELECTION FUNCTION *if for all* $W_{\text{new}} \subseteq W$ *(the "new worlds" to be added to A) and* $\Xi_{\text{new}} \subseteq \Omega$ *(the new attention state of the agent):*

- $f(W_{\text{new}}, \Xi_{\text{new}}) \subseteq W_{\text{new}}$ *(the function selects which elements of* W_{new} *are preferred), and*
- *for each* $w, v \in W_{\text{new}}$ *such that* $V_w \restriction \Xi_{\text{new}} = V_v \restriction \Xi_{\text{new}}$ *then either* $w, v \in f(W_{\text{new}}, \Xi_{\text{new}})$ *or* $w, v \in W_{\text{new}} \setminus f(W_{\text{new}}, \Xi_{\text{new}})$.

We need only require that $\mathfrak{B}$ be such a function, and our agent will always have belief-attention consistent beliefs after a pure attention update.

A final worry may occur to the reader: do such functions always exist? Might there not be bizarre states the agent could reach, in which no possible spontaneous belief will be attention-consistent?

LEMMA 3.13: Extreme personalities. *The following two functions are awareness-relative selection functions (i.e., they generate attention-consistent beliefs from all possible states):*

- *For each* W_{new} *and* Ξ_{new}, $f(W_{\text{new}}, \Xi_{\text{new}}) = W_{\text{new}}$ *(a* CREDULOUS *agent);*
- *For each* W_{new} *and* Ξ_{new}, $f(W_{\text{new}}, \Xi_{\text{new}}) = \emptyset$ *(a* CLOSE-MINDED *agent).*

Now we have the formal vocabulary we needed to fully define an agent's personality.

DEFINITION 3.14: Personality (definitive). *Fix a model* $M = \langle W, \Omega, V \rangle$. *An agent's* PERSONALITY *is a pair* $\Pi = \langle \preceq, \mathfrak{B} \rangle$ *such that*

- $\preceq$ *is a weak order on W; and*
- $\mathfrak{B}$: $\mathscr{P}(W) \times \mathscr{P}(\Omega) \to \mathscr{P}(W)$ *is an awareness-relative selection function.*

3 · *Properties*

The most important property of this system is that the dynamic formulation provides the same support relation as the static version of the previous chapter.

THEOREM 3.15: Equivalence of static and dynamic support relations. *Let M be a model of reality,* Π *an agent's personality, and* σ *a* Π*-consistent doxastic state. Then for all formulae* $\varphi \in \mathscr{L}_A$ *of the language of awareness:*

$$\sigma \Vdash^M \varphi \text{ iff } M, \sigma \models \varphi.$$

This is important not because it is unexpected but because it is reassuring: despite all the complications of the dynamic updates, we have not lost the simple idea that a state supports observations like "an agent in that state implicitly believes that φ but is unaware of it".

Here is another property which is in a sense the reason for including the assumption set in the model in the first place. Let $\sigma = \langle A, B, \Xi \rangle$ be an agent's state and let $\sigma[\varphi]_{\mathrm{n}} = \langle A', B', \Xi' \rangle$. Then $B = A \cap B'$. In words: whatever beliefs the agent had before an awareness update (B), she still holds them after the update *conditional on her original assumptions* ($A \cap B'$). Beliefs are defeasible under awareness updates: if I assume p and believe q then after becoming aware of the possibility that $\neg p$ I may no longer believe q. But I still believe *if p then q*: if my previous assumption turns out to be correct, my beliefs (which were conditional on that assumption) will also be correct.[7]

Some other properties relate more directly to the complications. For instance, if an agent has one of the extreme personalities given in Definition 3.13, her pure attention updates commute. Formally, for such an agent, $\sigma[\varphi]_{\mathrm{n}}[\psi]_{\mathrm{n}} = \sigma[\psi]_{\mathrm{n}}[\varphi]_{\mathrm{n}}$. This holds because of the way we select new worlds for the A set: they come from the ordering $\preceq$, which delivers the same results regardless of the order the worlds are selected in. The restriction to extreme personalities is needed, though, because changing the order of updates changes the particular sets of new worlds delivered to the spontaneous belief function $\mathfrak{B}$.[8] That is, any agent with any personality will get the same A set from updating with $[\varphi]_{\mathrm{n}}[\psi]_{\mathrm{n}}$ or with $[\psi]_{\mathrm{n}}[\varphi]_{\mathrm{n}}$, but their belief set B may differ on how it treats the new worlds unless their personality is suitably structured.

The following theorem is trivial to prove, given the way we have defined $[\,\cdot\,]$, but it illustrates the connection between our system and Veltman's:

Theorem 3.16: Updates under full awareness. *Fix $M = \langle W, \Omega, V \rangle$ and Π as before, and let $\sigma = \langle A, B, \Xi \rangle$ be a Π-consistent state. Then for all $\varphi \in \mathscr{L}^{\Xi,\Diamond}$:*

$$\sigma[\varphi]^M = \sigma[\varphi]^M_{\mathrm{b}}.$$

Equivalently, the above equality holds for all $\varphi \in \mathscr{L}^{\Omega,\Diamond}$ such that $M, \sigma \models N\varphi$: the formulae the agent attends to.

In particular, let σ be a Π-consistent state such that $\Xi = \Omega$ (the agent attends to all proposition letters in the model). Then the theorem applies for every formula in $\mathscr{L}^{\Diamond}$: no update can possibly change the awareness component of the agent's state.

This theorem tells us how we should interpret Veltman's model (recall that $[\,\cdot\,]_{\mathrm{b}}$ is his update, applied to the belief set). I have said already that pure belief updates do not occur outside our models, but a full conversational update with φ *has the same effect* as a pure belief update if the agent is already attending

[7] Note that the agent cannot necessarily *express* their assumption, even in the enriched language after the awareness update: it may have structure that requires awareness of yet more atomic propositions to describe.

[8] This restriction is sufficient but not necessary; I don't see any clear characterisation of the set of spontaneous belief functions that behave nicely here though.

to φ. Veltman's update should be thought of as the updates taking place between conversational participants who are already attending to every relevant propositional constant, so that the attention 'dimension' of their cognitive states holds constant throughout the conversation.

Since everything except the clause for *might* in Veltman's system is classical, for formulae without *might* it (and thus pure belief updates in my system) has the property that conjunction is equivalent to sequencing: $\sigma[\varphi]_b[\psi]_b = \sigma[\varphi \wedge \psi]$ (if φ and ψ do not contain *might*). This does not hold, however, for updates with awareness. Investigating the reasons why will raise some interesting questions about the interpretation of these models.

3.1 · *Sequencing and reinterpretation*

Pure belief updates with non-modal formulae genuinely accumulate information, in the sense that *how* the information was gained (the order of presentation and the syntactic form) is entirely forgotten by the updated state. This leads to some equivalences, among which that sequencing and conjunction are interchangeable. If φ and ψ are purely propositional then:

$$\sigma[\varphi]_b[\psi]_b = \sigma[\varphi \wedge \psi]_b = \sigma[\psi \wedge \varphi]_b = \sigma[\psi]_b[\varphi]_b.$$

This is no longer the case when we move to updates with awareness.

$$\begin{aligned}\sigma[\varphi][\psi] &=_d \sigma[\varphi]_n[\varphi]_b[\psi]_n[\psi]_b \\ \sigma[\varphi \wedge \psi] &=_d \sigma[\varphi \wedge \psi]_n[\varphi \wedge \psi]_b = \sigma[\varphi \wedge \psi]_n[\varphi]_b[\psi]_b\end{aligned}$$

It does not even hold (still for purely propositional formulae) that $\sigma[\varphi \wedge \psi]_n = \sigma[\varphi]_n[\psi]_n$, since the agent may form different spontaneous beliefs (according to the function $\mathfrak{B}$) depending on whether the updates come separately or are combined.[9] However if the agent's personality is CREDULOUS (all newly entertained worlds are held possible; see Lemma 3.13) then for purely propositional formulae, $\sigma[\varphi \wedge \psi]_n = \sigma[\varphi]_n[\psi]_n$. Still the two full updates differ:

$$\sigma[\varphi][\psi] = \sigma[\varphi]_n[\varphi]_b[\psi]_n[\psi]_b \tag{3.1}$$

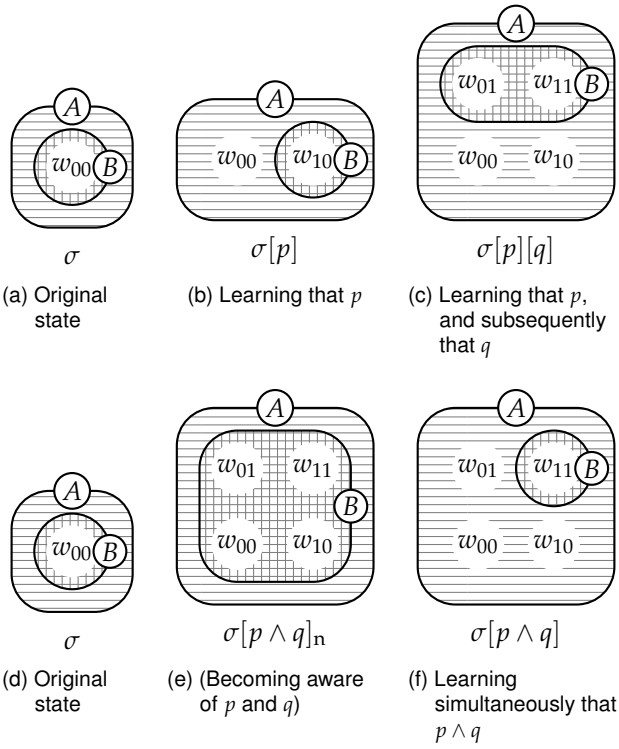
$$\sigma[\varphi \wedge \psi] = \sigma[\varphi]_n[\psi]_n[\varphi]_b[\psi]_b \qquad \text{(for credulous agents)} \tag{3.2}$$

This difference is not just a matter of notation: in (3.2) the belief update with φ takes place after the agent becomes aware of ψ, while in (3.1) the update with φ happens before the agent becomes aware of ψ. This can have serious

[9] If this property is undesirable for a particular application, it can be avoided by constraining the belief formation function: for $U_1, U_2, U_3 \subseteq W$ and $\Xi_1, \Xi_2, \Xi_3 \subseteq \Omega$, if $U_3 = U_1 \cup U_2$ and $\Xi_3 = \Xi_1 \cup \Xi_2$, then $\mathfrak{B}(U_3, \Xi_3) = \mathfrak{B}(U_1, \Xi_1) \cup \mathfrak{B}(U_2, \Xi_2)$. Both the extreme personalities listed above obey this constraint.

consequences: $\sigma[\varphi \wedge \psi] \models X\varphi$ holds for all states σ in all models ($\varphi \wedge \psi \models X\varphi$, as one would expect), but in general $\sigma[\varphi][\psi]$ need not support $X\varphi$.

To see why, and why this makes sense, let us take an example.



(a) Original state

(b) Learning that p

(c) Learning that p, and subsequently that q

(d) Original state

(e) (Becoming aware of p and q)

(f) Learning simultaneously that $p \wedge q$

Figure 3.3: Surprising attention dynamics: learning that q can overturn previously learned information that p. The top row corresponds to equation (3.1) on the previous page, the bottom row corresponds to (3.2). Worlds are labelled with pq-valuations, for example at w_{01} p is false and q is true. The agent is credulous, so awareness updates add all new worlds to the belief state. In the state σ she is aware of neither p nor q; her beliefs are entirely governed by her assumptions.

The agent in Figure 3.3 begins in a state of total unawareness. The awareness update with p (in the top row of the figure) adds w_{10} to her set A, while the belief update removes w_{00} from her B set; that is, learning that p overturns her assumption that $\neg p$. However, the belief update only eliminates worlds that she entertains; the awareness update $[q]_{\mathrm{n}}$ adds new worlds to her A set (w_{01}

and w_{11}), and since these worlds were not 'visible' when she updated with (the belief that) p, they are unaffected by that update. In the bottom row of the figure (the concurrent update), the awareness update with $p \wedge q$ already adds all four worlds to her assumption set; when the information content of "$p \wedge q$" is processed (the belief update), the world w_{01} does get removed, in contrast to the consecutive update.

In other words, attention updates introduce a very strong source of non-monotonicity into the system: an attention update can overturn nearly any belief whatsoever.[10]

This is the technical side of the question, but what about the realism of such behaviour? Well, suppose that p stands for "Airline travel has a smaller carbon footprint than train travel", and q for "The study showing that p was funded by Boeing". The sequence of updates can then easily be read as a dialogue: A says "Hey, this study is fascinating: p!" while B responds with "But q!" (Our agent might represent yet a third conversational participant, or she can be A if the first update $[p]$ represents her reading the study.)

Being simplified, this example is somewhat extreme; in particular, the reader may be suspicious of the fact that the interpretation of q makes direct reference to p. The problem is more general than such artificial examples, however. If state σ supports $X\varphi$ then an agent may consider herself justified in asserting φ.[11] How her assertion should be understood, however, is rather as $\zeta \rightarrow \varphi$, where ζ lists the agent's assumptions. In particular, she must assume many things about the sources of her own knowledge: that her senses function reliably, that the information reported to her is trustworthy, and so on. In the dialogue above, the hedged utterance of p should rather be "*According to this study*, p" (and indeed, for the simplistic example this is more natural); in general, though, we don't hedge all our assertions according to where the information we are reporting came from.

In Chapter 6 I will describe an attempt to distinguish along these lines between 'direct' (reliable) information and 'indirect' (and potentially unreliable) conclusions from that information. Rather substantial changes to the setting are needed in order to achieve this, however, and the resulting distinction strikes me as perhaps too idealistic. In reality, people in conversation *do* report unreliable information as if it were entirely trustworthy, and do sometimes

[10] "Nearly" because the implicit belief hedged by assumption is always preserved: if ζ exactly expresses the agent's assumptions before the update and she believes φ before the update, then after the update she is guaranteed to believe $\zeta \rightarrow \varphi$. The "exactly" is important, however: ζ must characterise her assumptions in the sense that a world is in A if and only if (not just "only if") it satisfies ζ. There may not always exist such a formula, and even when there does it very likely will not be found in the agent's language of awareness, even after the update.

[11] Loosely speaking; more precisely, the criterion for assertion should probably be something like "believes to know" rather than simply "believes" (of course explicitly in both cases). I return to this distinction in Chapter 4.

have to be confronted with the possibility that they have done so.

That said, however, these occasions are relatively infrequent in ordinary conversation. It is easy to see why that should be so: every time an awareness update overturns all previous beliefs, our agents have to go about *re-investigating* all the issues they thought were already settled. Three facts intervene in reality to prevent this.

The first is that such reinvestigation can be done 'internally': if I remember the utterances you have made as well as their information content, I can simply reapply all the updates corresponding to those utterances to my information state 'in my head'. Of course I should take some care in doing so: in the example above, the earlier update ("Airline travel has a smaller carbon footprint than train travel") should *not* be 'reapplied' in light of the new awareness (that, after all, is the point of the example!).

It is essential to the realism of the example, moreover, that the utterances p and q come from *different* agents. If q really undercuts the assertion that p, an agent who was already aware of both could not assert both. In this example, both the hearer and the agent asserting p are surprised by the possibility that q; that means that if more surprising assertions are needed, we need more agents! In other words, the second fact restricting the effect of non-monotonic updates of awareness is that no agent can surprise themselves.

The third reason why this non-monotonicity does not provide serious problems in real conversation is that we take pains to avoid it. If we have any suspicion that there are significant details that our conversational partners are unaware of, these are the first things we will want to draw their attention to. This is pragmatic 'good behaviour' exactly because surprises produce non-monotonic updates: our efforts at gaining and giving information will be wasted if the information is overturned by an awareness update, so we had better give the awareness update *first* and only afterwards the information.

3.2 · *Multiple information states*

Let us return for a moment to the observation that more surprises need more agents. The reason is that an utterance should be believed by the agent asserting it, at least at the moment of utterance. We need at least two information sources for the carbon footprint example, because the assertions that p and that q are partially in conflict: if the agent asserting p had been aware of q, he would not have made the assertion. Especially for cases where modelling credulous agents makes sense (where their assumptions seem genuinely unjustified and likely to be given up if challenged), each new atomic formula needs a separate advocate to introduce it. In the presence of multiple agents all potentially suffering from unawareness, reasoning about information becomes quite tricky. Here is an example, still rather small but already quite complex.

Example 3.17: Switches, and knowledge thereof. *The electrical wiring of a houseboat where I lived for a couple of years (this is a true story) had been put in by the man who built the houseboat, who was certainly no electrician. Particularly frustrating was the light for the entrance and stairwell: some days the switch next to the door worked, and other days it had no effect at all, apparently at random. It turned out that another flatmate suffered the same frustration, but with a* different switch *for the same light, placed at the head of the stairs. Instead of being wired as a two-way pair (as is usual for stairway lighting), these switches were in series with the bulb: if both were on the light was also on, but if either switch was turned off the light was off and the other switch would have no effect. My switch would work only if he had happened to leave his on, and vice versa.*

Both my flatmate and I were unaware of each other's switches; in propositional terms, we were unaware of the proposition letters representing whether those switches were on or off. We each made the entirely unconscious assumption that the switch of the other was off (or, stated positively, that the circuit from our own switch to the lightbulb was complete); on such an assumption, the failure of the light to turn on with the switch is evidence that the switch is broken (and the intermittent failure of the fault rather infuriatingly precludes tracking down the source of the problem). We would agree that there is something wrong with the light; by a quirk of natural language (not reproduced in my formal system) we could even agree that "The switch doesn't work" without realising that the referents of our respective referring expressions do not coincide.

If we had called in an electrician to sort the problem out, it would matter very much how we reported our beliefs to him. Suppose both switches happen to be off; I test mine and the light does not go on, and my flatmate tests his and likewise the light does not go on. If we both report to the electrician "The light is broken", he will form entirely false beliefs. On the other hand, if I report "If I turn my switch (downstairs) on, the light does not go on" (from which I had concluded that the light was broken) and my flatmate reports the same for the upstairs switch, the electrician will learn that there are two switches. He might then ask me exactly what the problem is; when I tell him that it *should* be the case that the light is on whenever my switch is on, he will know what is incorrect about my knowledge (namely that in fact the state of the light depends also on a second switch) and that he doesn't have to do any rewiring to fix the problem. From my own perspective the two statements are logically equivalent: given my assumptions, the light is broken just if it does not turn on when my switch goes on. But from a perspective of greater awareness, the two statements diverge; if I want to give the best information to the electrician, I had better give him the causally prior information rather than the conclusions I

have drawn from it.[12]

All of these complications can be dismissed when the utterances come from a single agent. On the assumption that an assertion of φ is only justified in state σ if $\sigma \models X\varphi$ (the agent explicitly believes that φ), if a single agent asserts φ followed by ψ then her state σ must support both φ and ψ, and so the *intended* force of the update would be not $[\varphi][\psi]$ but the stronger $[\varphi \wedge \psi]$. It would be a mistake, however, to model it directly as such. The actual effect on a listening agent is that first φ is processed and accepted, then ψ is processed and accepted. What may very well happen, though, is that if ψ brings new possibilities to the listener's awareness, she may recall the assertion of φ and *reapply* the belief update $[\varphi]_b$ to the new worlds. Then again, she need not do so: if she disagrees with some part of the later utterance she might decide that the earlier utterance is similarly suspect. (For more on rejecting updates see Section 4.2.)

A: I am a vegetarian.
B: [*Thinks: Then I won't offer you steak for dinner.*]
A: And of course, cows are a kind of vegetable.
B: No they aren't! And if that's what you believe, you aren't a real vegetarian either!

This kind of reassessment has to be driven by rather complicated processes that lie outside the scope of a formal theory; in particular, if Abel asserts φ then Ben brings up a possibility p that Charlie (who has been listening) was unaware of, poor Charlie has to figure out *somehow* if Abel had p in mind or not when asserting φ. The updates I have defined take a severely conservative strategy: they assume the worst (that is, the least awareness) of everybody, so that Charlie would have to ask Abel, "Do you still think that φ?"

Let us turn aside from these complications, and see some simpler possibilities that the system provides.

4 · *Applications*

4.1 · *Pragmatics of might*

The semantics I have given for *might* is exactly the same as Veltman's, however because it is embedded in an extended awareness account we can explain a

[12] Anyone who has worked in IT support will recognise this problem. A computer user will say something like "I didn't really do anything different, but yesterday it worked and today it doesn't." Careful questioning may uncover the fact that yesterday's work was on a different computer, or using different applications to do the same job, or with different versions of the same files (stored in different places) and so on and so forth. Unawareness of these possibilities (and the potential impact they can have) leaves novice computer users with a host of assumptions which must be painstakingly overcome before the root of the problem can be identified. This is why telephone helpdesk services often use a checklist that starts at the very beginning: check the computer is plugged in and turned on, the screen and keyboard are connected, ...

very natural usage of *might* which remains outside Veltman's account: drawing attention to possibilities.

This property is not limited to *might*: *any* mention of an atomic formula p makes the agent aware of p as a possibility.[13] What is special about *might* is that it achieves this end *efficiently*: it needs minimal epistemic grounding to be assertable, and (as Veltman realised), its *information content* is entirely negligible. It is a particularly *safe* way to draw attention: where saying p directly risks giving unreliable information, *might* p makes sure the hearer is aware of the possibility but leaves them to make up their own mind.

Outside of giving a noncommittal answer to a question, this is perhaps *the* most natural way to use *might*. It is only an artifact of the monotonicity properties of theories without unawareness that they cannot formally allow it to play this role: *coming to believe* that *might* φ is a non-monotonic update of a kind forbidden by (most) such theories. I don't even think that this interpretation of *might* is necessarily particularly new; what is new is that it is placed on a formal footing. Many people (perhaps Veltman among them) would agree that *might* is used to raise possibilities, but would leave that usage to the informal side of pragmatics: the formal theories describe instead what goes on against a stable background of attention to possibilities.

This leaves *might* in a rather odd position, however: it has the right conditions of acceptability, but it is hard to see why any speaker would ever want to announce a statement with *might*. I will illustrate this observation with respect to Veltman's system, but it applies equally to any model in which *might* expresses something about the speaker's information rather than information about the actual state of affairs.[14]

The most natural reason for an agent to make an announcement in the update setting is to induce a change in her conversational partner's state. That is certainly the main motivation for plain (non-modal) assertions: I assert "p" to induce you to come to believe that p. However this cannot be the motivation for announcing a *might* statement: since *might* is a test, it either produces no change in the hearer's belief state or takes her to the inconsistent state (which surely cannot be the intended effect).

We can get an idea of what Veltman's *might* could be used for by looking one step further in the dialogue. The hearer's response if her state already supports *might* φ is simply to accept; in that case neither information state gets

[13] It also makes the agent aware of $\neg p$. That is, this account still does not explain why drawing attention to p has such a different flavour from drawing attention to $\neg p$. I suspect that a proper account would start with a distinction between "being aware of" and "attending to" which I am deliberately eliding: we divide our attention unequally among the possibilities we are (equally) aware of.

[14] These remarks apply only to a top-level 'assertion' with the form *might* φ; Veltman's account also treats embedded *might*, where it is easier to come up with sensible speaker motivations.

changed and nothing is achieved. But if her state already supports $\neg\varphi$, then she must *reject* the update to avoid arriving in the inconsistent state. This rejection gives the utterer of "*might* φ" pragmatic information (that her state supports $\neg\varphi$), but a cooperative hearer might be expected to go further and make a follow-up assertion explaining why she believes that φ is not a possibility. The agent who utters "*might* φ" can be seen as probing for exactly this; the usage is somewhat analogous to a *question*, rather than an assertion, in that its pragmatic purpose is to prompt the hearer into giving information.

Veltman's pragmatics are still available for my agents, but I suspect they will make little use of them (and in real conversation, this reading of a *might* statement seems quite unusual unless the utterance comes with questioning intonation, in which case arguably the representation should include an explicit question operator). What looks far more familiar is the use of *might* to coordinate attention: to make sure that everyone is aware of the same set of live options. Since "*might* φ" mentions all the proposition letters in φ, the awareness update it triggers makes sure that the hearer entertains some φ possibilities. However, unlike a direct assertion, it does so without imposing any substantive update on the hearer. This explains why something as truth-conditionally weak as a possibility modal can still be usefully employed in conversation: its associated awareness update can affect the hearer's state far more extensively than its truth conditions would.[15] Chapter 5 gives a more precise statement of this observation, dealing explicitly with the conditions under which an awareness update can be considered 'a relevant utterance' or 'cooperative behaviour'.

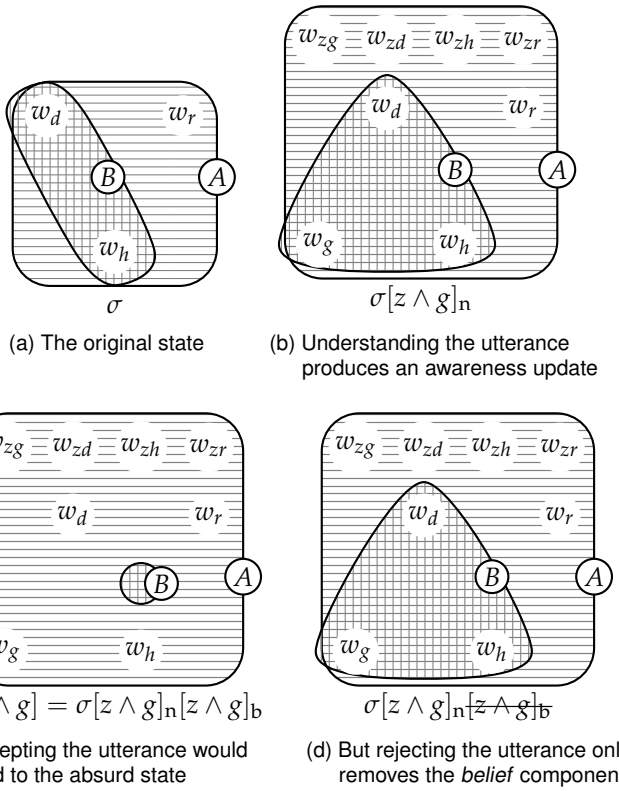
4.2 · *Rejecting updates*

Suppose I were to announce to you that I am the Lizard King. Your natural reaction would not, I hope, be to believe me. You would *reject* the update to your belief state (or to the common ground, if you prefer) that I propose by making such an assertion. (Analogously to the account given for Zeus

[15] This notion is not confined to *might*, but applies to any sentence with weak truth conditions. Even the fact that existentials introduce discourse referents almost looks like an awareness effect if you squint at it the right way (I owe this suggestion to Robert van Rooij). Disjunction under a possibility modal is another example. Hans Kamp, writing on free choice permission, gives a number of related examples that are "typically used in speech acts whose function it is to bring a certain number of possibilities to the attention of the audience. [...] In each of these cases the speaker's use of the disjunction can be interpreted as testifying to his indifference whether the possibilities he wants to bring to the hearer's attention satisfy the first or the second disjunct; and so it may be inferred that he wants to bring both kinds of possibilities to his attention" [Kam78, pp. 281–282]. Under a formal model of unawareness, of course, no such inference is necessary: a disjunction simply has the effect of drawing attention to both its disjuncts. I do not mean to suggest that awareness models will solve the problem of free choice permission, but merely to note that *intuitions* regarding attention to possibilities can be found at the informal edges of a wide range of more formal linguistic theories, which could perhaps be fruitfully extended by making these intuitions more formal.

above, your spontaneous belief would exclude the Lizard King possibility; then accepting the update would take you to the absurd state, so you must reject it.) Let σ be your original cognitive state, and φ my utterance; I have proposed $[\varphi]$ and you have rejected the update. In what cognitive state do you end up?

Certainly not $\sigma[\varphi]$ (the absurd state), or your rejection will be worth very little. But also not σ, for you cannot pretend that I have not made the assertion I have. I suggest that the correct answer is $[\varphi]_{\mathrm{n}}$: I proposed $[\varphi] = [\varphi]_{\mathrm{n}}[\varphi]_{\mathrm{b}}$, and while you can (and should) reject the *belief* component of the update, you are simply not capable of rejecting the *attention* component. Here is the picture for Olga, when her flatmate Ella tells her "It might be Zeus disguised as a goose."



(a) The original state

(b) Understanding the utterance produces an awareness update

(c) Accepting the utterance would lead to the absurd state

(d) But rejecting the utterance only removes the *belief* component

Figure 3.4: The update two-step. A pure awareness update takes the agent from (a) to (b); accepting the information component would take her to (c), while rejecting it leads to (d): the awareness update survives.

4.3 · *Dynamics of counterfactual conditionals*

The final application I will discuss is to a puzzle about counterfactual conditionals that has acquired the name 'Sobel sequences'. In a nutshell, counterfactuals exhibit ordering sensitivity which I argue is well explained by awareness dynamics. More than a straightforward application of the theory, this problem stands as a full-blown case study and the subject of the following chapter.

Chapter 4

Case study: Sobel sequences

> "Apparently there is no limit," Joe remarked. "Anything can be said in this place and it will be true and will have to be believed."
>
> Flann O'Brien, *The Third Policeman*

Since [Sta68] and [Lew73], a SIMILARITY-BASED account of counterfactual conditionals has been pretty much standard in philosophical logic. According to such a theory (in very broad brush-strokes), a counterfactual conditional as in (1) is true in a world w if in all those worlds where p holds *which are most similar to* w, q holds also.

(1) If p were the case, q would have been. [*Notation:* $p \mathrel{\Box\!\!\rightarrow} q$]
   a. If Sophie had gone to the New York Mets parade, she would have seen Pedro Martínez.
   b. If the US threw all its nuclear weapons into the sea tomorrow, there would be war.

Two significant problems exist with this account, which this chapter will address. The first concerns the notion of 'similarity': if this cannot be given a systematic foundation, then a semantics based on similarity cannot be truly explanatory. The second problem concerns the data: some combinations of counterfactual sentences ('Sobel sequences') show context-dependent dynamic behaviour which cannot be explained on the static account, and the most influential solutions for this problem have advocated giving up the similarity-based account almost entirely.

I shall take these problems in reverse order. My first aim is to explain the dynamic behaviour of Sobel sequences as a minimal addition to a similarity-based account, adding nothing but awareness dynamics to the picture. The explanation will provide quite strong constraints on the content of 'similarity', which fit in neatly with the most recent attempts to explicate 'similarity' in causal terms.

DEFINITION 4.1: Object language with counterfactuals $\mathscr{L}^{\Box\!\!\rightarrow}$. *The object language with counterfactuals adds a binary counterfactual conditional operator $\Box\!\!\rightarrow$ to the syntax of $\mathscr{L}^{\Diamond}$.*

Recall that $\mathscr{L}^{\Diamond}$ is the propositional language with *might*. As I said when introducing the first object languages, I prefer to leave the syntax relatively loosely constrained: $(\mathit{might}\, p) \Box\!\!\rightarrow (q \Box\!\!\rightarrow (\mathit{might}\, r))$ is strictly speaking a well-formed formula. Most nestings will not make sense with the semantics I give, but *might* in the consequent of a counterfactual is deliberately included.

1 · *Counterfactual dynamics: the problem*

Since I want to defer the exploration of just what 'similarity' should mean to Section 3, let me follow [Lew81] and refer to ORDERING SEMANTICS; indeed, I will largely follow his presentation of ordering semantics in that paper.

DEFINITION 4.2: Ordering semantics (after [Lew81]). *Let W be a finite set of worlds. An* ORDERING FRAME *for W associates with each world w a preorder $\leq_w$ on W (that is, a reflexive and transitive binary relation $\leq_w \subseteq W \times W$), satisfying the following condition:*

Centering: *For all $w, v \in W$: if $v \neq w$ then $w <_w v$.*[1]

We write $\mathcal{S}^W_{\leq}$ for such an ordering frame ($\mathcal{S}$ for "similarity"); it is a function from elements of W to orderings on W, but we write $\leq_w$ for the ordering associated with $w \in W$, instead of the more cumbersome $\mathcal{S}^W_{\leq}(w)$.

Given any preorder $\leq$ on W, the set of $\leq$-MINIMAL φ-WORLDS *is written* $\min_{\leq}(W\restriction\varphi)$ *and defined*

$$\{w \in W \,;\, w \models \varphi \wedge \neg\exists w' \in W : w' \models \varphi \wedge w' < w\}.$$

(The finite setting guarantees the well-behaved existence of such sets.) The set of CLOSEST φ-WORLDS TO w *is given by*

$$\min_{\leq_w} W\restriction\varphi.$$

The counterfactual $\varphi \Box\!\!\rightarrow \psi$ is true at world w if all closest φ-worlds to w are ψ-worlds:

$$w \models \varphi \Box\!\!\rightarrow \psi \textit{ iff } \forall w' \in \min_{\leq_w} W\restriction\varphi : w' \models \psi.$$

The ordering semantics stands in opposition particularly to a STRICT analysis (more standardly assumed for indicative conditionals), under which $p \Box\!\!\rightarrow q$ would be true if q holds in *all* the worlds where p does. (The set of possible worlds must be contextually restricted in some fashion, but according to the

[1] I assume Lewis's constraint "Universality": that all worlds are included in the field of $\leq_w$.

strict analysis this restriction is not directly related to the semantics of the counterfactual.) The strict analysis validates the inference pattern 'strengthening the antecedent': from $p \boxrightarrow q$ it follows that $p \wedge r \boxrightarrow q$. That counterfactual conditionals do not allow strengthening the antecedent was a strong motivation for the ordering semantics; here is an example, which we will see reappear in many variations throughout the rest of the chapter.[2]

(2) a. If Sophie had gone to the New York Mets parade, she would have seen Pedro Martínez.
b. But if Sophie had gone to the parade and been stuck behind a tall person, she would not have seen Pedro.

An ordering semantics allows these two counterfactuals to be *simultaneously* true at a world w, if the ordering relation is suitably chosen. However something more seems to be going on. [Fino1] credits Irene Heim with the observation that in reverse order this sequence no longer sounds felicitous:

(3) a. If Sophie had gone to the New York Mets parade and been stuck behind a tall person, she would not have seen Pedro Martínez.
b. #But if Sophie had gone to the parade, she would have seen Pedro.

NOTATION 4.3: Some informal terminology. *A* SOBEL SEQUENCE *is a pair of counterfactuals with the forms shown in (2) and (3). The* FORWARD SEQUENCE *is as in (2), the* REVERSE SEQUENCE *as in (3). Based on the number of distinct possibilities alluded to, I call (2-a) and (3-b)* SIMPLE, *and (2-b) and (3-a)* COMPLEX.

The Sophie Sobel sequence is by no means an isolated example — perhaps even the majority of 'natural' examples of the failure of strengthening the antecedent behave in this way. (It turns out, however, that the pattern is not universal; I will give examples below that share the same logical structure but for which both orderings are felicitous.) The challenge to the ordering semantics is clear: as a static theory it cannot represent the influence of ordering in discourse, so there is no chance for it to explain these data.

I will argue for a modular approach: the core of the ordering semantics should be left exactly as it is, but combined with a separate component dealing with dynamic effects; I want to argue that the context-change in these examples is properly modelled as an awareness update (and thus has nothing inherently to do with counterfactuals at all).

This account has three benefits: it lets us keep the familiar ordering semantics, with all its acknowledged benefits; since the mechanism of awareness

[2] I have the example from [Mos07]; [Gil07] has "If Sophie had gone to the parade, she would have seen Pedro dance." According to Wikipedia, Pedro Martínez was a pitcher for the New York Mets from 2005 until 2008 and is now a free agent; I don't know whether he typically danced at parades.

update is not associated with counterfactuals, we can reuse the same explanation for a large number of other constructions that pattern like Sobel sequences (Section 4); and the specific requirements of this account for the ordering semantics will throw some light on what 'similarity' should mean (Section 3).

First, though, I have to discuss an alternative explanation for the contrast between (2) and (3), which has recently grown in popularity. The reader more interested in my positive contribution is invited to skip ahead to Section 2.

1.1 · *Strict semantics and shifting context*

At first sight the alternative is extremely seductive, especially in its more elaborate formulations. [War81] is the earliest version I have seen, but [Low90; Fin01; Gil07; Wil08] (among others) all give variants of this proposal,[3] and I will cheerfully mix their terminology to suit myself. The version I give is very simple, and I must ask the reader to accept (or to check for themselves) that the bells and whistles added by the various accounts that I conflate will not affect my critique. Here, then, is the proposal in a nutshell; I call it the SHIFTING STRICT analysis.

We are to evaluate a counterfactual according to the strict semantics, against a contextually given set of worlds, the MODAL HORIZON of [Fin01]. A counterfactual $\varphi \boxdot\!\!\rightarrow \psi$ comes with what [Gil07] calls an ENTERTAINABILITY PRESUPPOSITION,[4] that φ be satisfiable within the modal horizon. If the presupposition is not met, accommodation adds some worlds to the modal horizon; the worlds added are the closest φ-worlds to the evaluation world, according to the same kind of similarity ordering as is needed for ordering semantics. After accommodation, the strict truth conditions are simply: $\varphi \boxdot\!\!\rightarrow \psi$ is true (at w) if every world in the modal horizon (of w) which supports φ also supports ψ.

It is easy to see that such an account explains the Sobel data. Entertainability presuppositions only *add* worlds to the modal horizon, so counterfactuals later in the discourse 'inherit' the possibilities introduced by earlier utterances (such as Sophie being stuck behind someone tall). More subtly, it can provide exactly the same predictions about *single* counterfactuals (uttered in the 'null context') as does the ordering semantics: if the modal horizon in the null context is

[3] Warmbrōd was concerned with the inference pattern of substitution of equivalent antecedents rather than strengthening the antecedent; his proposal suffers in readability, through no fault of its own, by predating modern notions of dynamic semantics. Williams is concerned with indicative, rather than counterfactual, conditionals; we will consider some of his data below, in extending the account beyond counterfactuals. The account I give is based most directly on the von Fintel analysis. My impression is this has been relatively influential, which I find somewhat surprising in light of the simple counterexamples that I will introduce below. Sarah Moss is a welcome voice of scepticism, and I will draw heavily on her —as yet unpublished— account [Mos07] in what follows.

[4] I don't think it is essential to these accounts that entertainability 'presuppositions' be presuppositions as usually conceived (in fact there are both systematic similarities and systematic differences). Using the term will help the clarity of my account, if I may do so without taking it too seriously.

suitably trivial (empty or containing only the evaluation world), the first update will add to it precisely those worlds from the ordering that would influence the truth conditions according to the ordering semantics.

1.2 · *Problems with the shifting strict analysis*

The way the shifting strict analysis incorporates single counterfactuals marks a profound difference in methodology from my own approach. Both analyses must acknowledge the fact that the ordering semantics provides extremely intuitive predictions for single counterfactuals (in the 'null context'). My approach is to *augment* the ordering semantics with a component dealing with context change; the predictions in the null context still come from the same core mechanism in the ordering semantics. The shifting strict approach, on the other hand, *supplants* the ordering semantics; the predictions in the null context come from a different mechanism, which is carefully adjusted to produce coinciding predictions. This produces an odd redundancy in the system: the strict semantics is for counterfactuals, but the shiftiness is for counterfactuals too; the theory in some sense contains intertwined both a static and a dynamic counterfactual semantics.

I think this methodological distinction already provides reason to prefer my analysis, if the two can be shown to perform equally well on the data: I prefer a 'modular' approach in which the dynamic behaviour of counterfactuals emerges from interaction between a counterfactual semantics and a dynamic theory of discourse, where the two are much more distinct and independent than the shifting strict analysis will allow. I will go further, though, and suggest that there would be good reason to reject the shifting strict analysis even if we had no better account of the dynamics.

I want to raise three problems, respectively observational, conceptual, and methodological: the shifting strict analysis mispredicts on some simple examples, close to the core of what it is designed to explain; it rejects without good reason the possibility of simultaneously considering different 'levels of counterfactuality'; and it fails to capture the generality of the Sobel forward-and-reverse pattern, by limiting its account to (at best) conditional sentences.

1.2.1 · *Prediction failures*

Each of the following sequences is predicted by the shifting strict analysis to be infelicitous, in the same way that the reverse Sobel sequence is. (I have (4-c) from [Mos07], where it is credited to John Hawthorne.) To my ear, at least, none of them bear out that prediction.

(4) a. (i) If Sophie had gone to the parade and not seen Pedro, she wouldn't have seen Pedro. $p \wedge \neg q \boxrightarrow \neg q$

(ii) But if Sophie had gone to the parade, she would have seen Pedro. $p \boxrightarrow q$

b. (i) If Sophie had gone to the parade and not seen Pedro, she would have been upset. $p \wedge \neg q \mathrel{\Box\!\!\rightarrow} s$
   (ii) But if Sophie had gone to the parade, she would have seen Pedro. $p \mathrel{\Box\!\!\rightarrow} q$

c. (i) If Sophie had gone to the parade and been shorter than she actually is, she would not have seen Pedro.
   (ii) But if Sophie had gone to the parade, she would have seen Pedro.

d. (i) If Sophie had gone to the parade and been eaten by a dinosaur, she wouldn't have seen Pedro. $p \wedge d \mathrel{\Box\!\!\rightarrow} \neg q$
   (ii) But if Sophie had gone to the parade, she would have seen Pedro. $p \mathrel{\Box\!\!\rightarrow} q$

Of course any theory can deal with a limited number of counterexamples: there is always the possibility that further wheels upon wheels may be added that will save the phenomena. How attractive such a strategy appears should depend on how central the counterexamples are for the class of phenomena to be explained; my impression of these examples is that they are paradigm cases of counterfactual use,[5] and that a semantics for counterfactuals should explain such data in its core rather than at the level of a corrective epicycle.

The four examples divide naturally into two classes. The first two involve entertaining the possibility that Sophie does not see Pedro, but (unlike the classic Sobel sequence) provide no reason to take this possibility seriously. The last two examples have a different flavour: they provide reasons why Sophie might not see Pedro, but the reasons themselves are not to be taken seriously.

All these examples *can* be modelled using orderings, but I want to make a stronger claim. In cases like the first two, where we have no reason to expect Sophie to not see Pedro, I only want to allow orderings in which she at least *might* see him (if she goes to the parade, naturally). Section 3 gives an account of where the orderings of ordering semantics *come from*, which fulfils this desideratum.

For the second two cases, it is enough for me that ordering semantics makes it possible that they be acceptable, since the shifting strict account does not. This leads us to the conceptual critique of the shifting strict analysis.

1.2.2 · *Simultaneity ruled out*

Under the ordering semantics the counterfactuals making up a Sobel pair can be *simultaneously* true. In the cases Heim noticed this doesn't seem right; the impression we have on reading the forward sequence is that in light of the complex utterance, the simple is no longer true. However examples such as (4-d) above should make us question whether this pattern is (as the shifting

[5] That the acceptability of (4-a) does not depend on its tautological status is shown by (4-b).

strict analysis would have it) truly universal. It seems perfectly reasonable to believe *simultaneously* that Sophie would not have been eaten by a dinosaur at the parade, and nonetheless that from inside a dinosaur she would not have been able to see Pedro.

We might think of this example in terms of embeddings of counterfactual contexts. At the factual level, Sophie doesn't go to the parade. In the first counterfactual context, she does go and sees Pedro — and is of course not interfered with by dinosaurs, which live only in a second counterfactual context, embedded in the first.

A counterfactual is most natural in a context where its antecedent is known to be false; we can agree at least on this, without taking a stand on precisely where this naturalness comes from (presupposition, Gricean inference, or whatever your favourite explanation may be). We can see this as a kind of concession: "I admit that in our current context φ is the case, but if we go to an embedded context where it isn't then. . . ." And this works equally well for higher levels: "I admit that in the embedded context where Sophie goes to the parade she is not eaten by a dinosaur, but if we go to yet further embedded context where she is then. . . ."

The odd thing about the shifting strict analysis is that it treats these two cases as fundamentally different. In the first case, what is 'conceded' is sacrosanct and cannot be affected by the counterfactual utterance; but in the second case, the concession is undermined by the effect of accommodating the entertainability presupposition. After "I agree that she didn't, but suppose that she had", "she didn't" is still supported; after "I agree that she wouldn't have, but suppose that she did", "she wouldn't have" is not.

Here is a similar, more extended example.

(5)
a. A: Suppose Sophie had gone to the parade yesterday.
b. B: She would have seen Pedro.
c. A: But suppose she was eaten by a dinosaur.
d. B: She wouldn't have been!
e. A: Sure. But suppose she was. Then she wouldn't have seen Pedro, right?
f. B: Alright.
g. A: So if she had gone to the parade, would she have seen Pedro?
h. B: Of course she would have.

I don't want to claim that "Suppose that φ were true; then ψ would be" always has a meaning identical to "If φ were true ψ would be", but the intuitive similarity should not be dismissed either. Intuitively we distinguish different nested counterfactual contexts in these examples, and so those distinctions should be available for our counterfactual semantics to work with.

1.2.3 · *Missing generalisation*

The final significant deficiency of the shifting strict analysis is that it has too narrow a focus and too specific a mechanism. The analysis posits particular features of the semantics of counterfactuals underpinning the Sobel pattern; if the same pattern occurs without involving counterfactuals at all, the shifting strict account has nothing to say about it.[6] Sarah Moss has collected a wide range of non-conditional examples showing the Sobel pattern (as in (6)), and gives an intuitive explanation of their common structure in an unpublished manuscript [Mos07]; her explanation requires no revision of standard counterfactual semantics.

(6) a. (i) A: My car is around the corner.
(ii) B: Cars get stolen all the time here in New York City.
b. (i) B: Cars get stolen all the time here in New York City.
(ii) A: ? My car is around the corner.

Her observation is that attention to possibilities affects the assertability of perfectly ordinary statements as well as counterfactuals. I discuss some more of her data in Section 4, which extends my own theory beyond counterfactuals, since I disagree slightly on its interpretation, but I am far more sympathetic to her mode of explanation than to that of the shifting strict analysis.

1.3 · *Desiderata for a replacement theory*

I have dwelt on the shifting strict analysis at some length, because its shortcomings outline a number of desiderata that must be achieved if any theory is to qualify as a potential replacement.

Most obviously, we need a dynamic account. The dynamics cannot be driven by conditional form (if we are to cover Moss's data); our theory should not predict that every sequence with the Sobel form has the Sobel pattern of acceptability (to cover the core counterexamples of (5)). We must incorporate enough of the ordering semantics to be able to represent nested levels of counterfactual supposition.

However, Heim's observation and the widespread acceptance of (something like) the shifting strict analysis points at another, rather more subtle, desideratum. Sobel-pattern pairs are easy to think up, but pairs that are felicitous when reversed are quite a bit thinner on the ground. The shifting strict analysis rules such pairs out entirely, and falsely; but our theory should, ideally, have

[6] In fact not quite every account that I gather under the heading "the shifting strict analysis" is concerned solely with counterfactuals. The account of [Wil08] applies to indicative conditionals rather than counterfactuals, and is in some respects closer to my own views. It still ties the mechanism to the specific semantics of conditionals, though (entertainability presuppositions are triggered only by conditional antecedents); I am arguing for a general framework of awareness-sensitive semantics which applies equally to conditionals and to any other construction.

something to say about the *relative frequency* (and naturalness) of the two kinds of examples.

I will now give a theory that fulfils these desiderata.

2 · Orderings with awareness

The idea is simple in the extreme. We will simply add an ordering frame to the dynamic awareness models of the previous chapter, as a component of the model of reality; a counterfactual conditional will be evaluated according to the ordering semantics but on the awareness state of the agent, which may exclude a number of possibilities due to the agent's assumptions. Changes in awareness will update this state as I described in the previous chapter; it is these dynamics, rather than anything particular to counterfactuals, which will give rise to the pattern noted by Heim.

The ordering semantics I gave in Definition 4.2 was for a single world, so our first task is to lift it to an information state. It might seem that a state should support the counterfactual if all worlds in the state do so, but the possibility of *might* occurring in the consequent requires a more careful approach. If I don't know whether it is raining in Whitechapel or not, I can perfectly acceptably tell you "If you were in Whitechapel you might have been rained on"; presumably this is true even though one of my epistemic alternatives (where it is not raining in Whitechapel) does not itself support the counterfactual. [Velo5] has a similar starting point (although the semantics end up quite different). Veltman defines the result of updating an information state with $\varphi \mathrel{\Box\!\!\rightarrow}$ (something like "Suppose that φ"); the new state collects the nearest φ-neighbours of all the worlds in the state (call that set C, for "counterfactual possibilities"), and the consequent of the conditional is tested against this set as a whole.

This means that a counterfactual cannot carry information: like any test, it either leaves the information state unchanged or takes it to the absurd state. In Section 4.2 I give an alternative definition which allows counterfactuals to be informative ([Velo5] also gives both a test and an update version of his semantics), and some reasons to suppose that we need both definitions.

Definition 4.4: Ordering semantics for attention models (counterfactuals as tests). *Let $M = \langle W, \Omega, V, \mathcal{S}^W_{\leq} \rangle$ be a model of reality augmented with an ordering frame, and $\sigma = \langle A, B, \Xi \rangle$ a state. We treat a counterfactual as a test:*[7]

$$\sigma[\varphi \mathrel{\Box\!\!\rightarrow} \psi]^M_{\mathrm{b}} =_d \begin{cases} \sigma & \text{if } \langle A, C, \Xi \rangle \models^M_{\mathrm{b}} \psi, \\ \langle A, \varnothing, \Xi \rangle & \text{otherwise,} \end{cases}$$

$$\text{where } C = \bigcup_{w \in B} \min_{\leq w} A \restriction \varphi$$

[7] This definition formally allows for nesting counterfactuals; I make no claim for its appropriateness in such cases, however.

Note that this definition gives only the *pure belief* update $[\,\cdot\,]_{\mathrm{b}}$. The key to the proposal is that the update $[\varphi \mathrel{\Box\!\!\rightarrow} \psi]$ is, as in the previous chapter, analysed in two steps: $[\varphi \mathrel{\Box\!\!\rightarrow} \psi]_{\mathrm{n}}[\varphi \mathrel{\Box\!\!\rightarrow} \psi]_{\mathrm{b}}$. The first step works just as any other attention update: it draws attention to atomic formulae mentioned in the counterfactual. It is only the second step that needs an explicit definition for the counterfactual connective.

For purely propositional φ and ψ the definition I give comes to the same as checking whether each world in B individually satisfies $\varphi \mathrel{\Box\!\!\rightarrow} \psi$ with the standard world-based ordering semantics for the counterfactual, however when *might* appears in the consequent the results can differ.

(7) The barroom tough Big Joe has invited you outside for fisticuffs, an invitation you have politely declined. You are not sure if Big Joe is drunk or not; sober he is unstoppable, but drunk he is as likely to knock himself out as to down his opponent. Describing the events the following day, you say:

a. "If I had taken the fight, I might have won."

The semantics given above correctly predicts (7-a) to be true in this context; the world-based formulation with universal quantification, on the other hand, would make it false. (There is an epistemic alternative, namely the one in which Joe is sober, for which all closest fight-alternatives have you losing.)[8]

Using such a semantics any Sobel-pattern pair of counterfactuals can be made simultaneously true, or incompatible, as required (the ordering semantics is all that is needed). However in order for the forward and reverse judgements to come out right under the awareness update, we need a particular structure on the ordering relation. Let me give the representation of Sophie's visit to the parade first as a 'Just So story', and show how it matches the data. After that I can say what the account requires in general, and how orderings satisfying those constraints might be generated. Figure 4.1 on the facing page gives the worlds and similarity relation we will need.

We assume for the moment that the agent holds only w_0 possible ($B = \{w_0\}$), so we are only concerned with the ordering for that world. Most of the details of the ordering should be uncontroversial; the only element that is both necessary for my account and potentially unexpected is the equisimilarity of w_5 and w_6. In fact it is only required for my account that w_5 be not more similar to w_0 than

[8] Note that this discussion is independent of the well-known debate on whether the law of conditional excluded middle holds *at a particular world*. If it does, then there are three relevant epistemic alternatives: one in which Joe is sober, and two in which he is drunk (differing only in who wins in the closest alternative in which you fight). I prefer a formulation in which the uncertainty about the outcome of the fight is *metaphysical* rather than *epistemic*, which I will argue for in Section 3. In either case, however, the lifting to information states provides counterexamples to conditional excluded middle — just as an information state may record neither certainty that p nor that $\neg p$.

	w_0	w_1	w_2	w_3	w_4	w_5	w_6	w_7
p	0	0	0	0	1	1	1	1
q	0	0	1	1	0	0	1	1
r	0	1	0	1	0	1	0	1

$$w_0 <_{w_0} \{w_5, w_6\} <_{w_0} \{w_4, w_7\} <_{w_0} \{w_1, w_2, w_3\}$$

Figure 4.1: Worlds for a Just So story about Sophie. The proposition p represents "Sophie went to the parade", q is "Sophie saw Pedro" and r is "Sophie got stuck behind someone tall". The actual world is w_0, and we assume the agent knows this; thus only the similarity relation for w_0 (shown below the table) will be needed.

w_6 is; equisimilarity is simply the most natural way to fulfil this condition (and will be implied by the causal semantics given in Section 3).

We need to give the agent's attention ordering $\preceq$ (used to specify which worlds 'spring to mind' depending on which atomic formulae she attends to). Here is the ordering we need:

$$w_0 \preceq \{w_2, w_4, w_6\} \preceq \{w_1, w_3, w_5, w_7\}$$

It encodes a default assumption that r is false (that Sophie is not stuck behind anyone tall). This Just So element, too, must be justified at a later date.

Knowing $\preceq$ we can generate A (the worlds entertained) from Ξ (the propositions attended to). If the agent attends only to $\{p, q\}$, her assumption set A is $\{w_0, w_2, w_4, w_6\}$ (that is, she assumes that r is false). This is the sort of assumption-by-omission that we expect when agents fail to attend to specific possibilities. Then the worlds in her cognitive state can be pictured as:

$$w_0 <_{w_0} w_6 <_{w_0} w_4 <_{w_0} w_2.$$

Assuming that w_0 is the only world in B, this clearly supports $B(p \boxdot\!\!\rightarrow q)$. Call this state σ_1.

If the agent becomes aware of r, however, her state expands to include all eight worlds:

$$w_0 <_{w_0} \{w_5, w_6\} <_{w_0} \{w_4, w_7\} <_{w_0} \{w_1, w_2, w_3\}$$

Call this state σ_2. Now σ_2 no longer supports $B(p \boxdot\!\!\rightarrow q)$, although it does support $B((p \wedge r) \boxdot\!\!\rightarrow \neg q)$. And what could trigger such an expansion? The simplest possibility is, hearing "$p \wedge r \boxdot\!\!\rightarrow \neg q$". One feature of forward Sobel

sequences is how naturally they read as an argument between two participants, instead of a pair of utterances coming from one agent.

Let σ_0 be a state of no awareness, where the agent entertains only w_0. Then the schematic picture of the possible dynamics looks as in Figure 4.2.

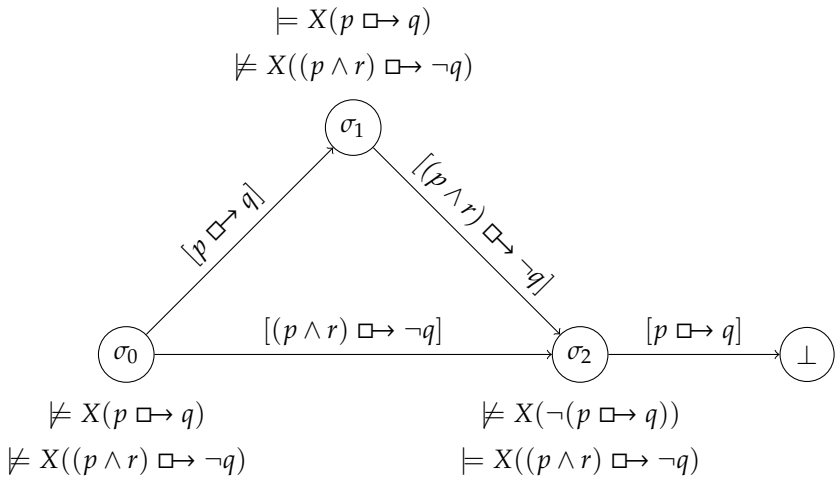


Figure 4.2: Awareness states and possible updates. State σ_0 is the state of no awareness; recall that X is the *explicit* belief operator. In all states only w_0 is in the belief set B. The belief updates are all tests, so the states only change because of the awareness component: at σ_0 and σ_1 the agent does not attend to r, while at σ_2 she does. The state labelled "$\bot$" has an empty (absurd; inconsistent) belief set.

Note that the shifting strict semantics would produce the same transitions. That is, with the attention dynamics as I have specified, we get exactly the same predictions as from the shifting strict semantics. However, since this is only one possible parameter setting, the present account captures a range of phenomena that the shifting strict semantics cannot.

2.1 · *Dinosaurs and tautologies*

We capture the counterexamples I gave to the shifting strict semantics (in Section 1.2.1) without major difficulties, on natural assumptions about the orderings involved. In fact, nothing extra is needed to represent the first two examples.[9] For (4-d) I have to do a little more work: I have to convince you that

[9]Unless something truly pathological happens when the agent becomes aware of the possibility that Sophie gets upset; I ignore the possibility.

it is *distinctly unlikely* that Sophie would get eaten by a dinosaur if she went to the parade. Recall that to get the standard Sobel sequences right, I have to assume that w_5 and w_6 are equisimilar to the actual world. This is defensible, on a suitable notion of 'similarity' (see Section 3), however the same cannot be said for the corresponding worlds (with and without dinosaurs at the parade) of (4-d)! The same holds (in less spectacular fashion) for (4-c): Sophie being shorter than she actually is need not be taken as seriously as her being *exactly as tall* as she actually is (again see Section 3).

2.2 · *Influence of factual information*

(8) a. A: If I had taken the fight and Big Joe was drunk, I might have won.
b. B: Big Joe wasn't drunk.
c. A: Ah. Then if I had taken the fight I would have lost.

It is perhaps not surprising that the system makes the correct prediction for this example , but I mention it in particular as an argument for modular system design. Factual assertion ("Big Joe wasn't drunk") removes worlds from the belief set, while the counterfactual test takes the belief set as its starting point when collecting most similar φ-possibilities ("If I had taken the first…"). The interplay of these orthogonal systems produces exactly the effect we want, without needing any special rules for the combination.[10]

2.3 · *Kernel of the account*

So far I have given only specific examples of the awareness account in action. Here are the general features. Suppose the Sobel sequence has the form $p \boxrightarrow q; p \wedge r \boxrightarrow \neg q$. What is required to make this account work?

For Sobel sequences on the Heim pattern (felicitous forward but infelicitous backward) we need three features:

1. Full awareness still justifies $(p \wedge \neg r) \boxrightarrow q$;
2. Full awareness justifies $p \boxrightarrow \textit{might}\, r$;
3. Unawareness of r produces a default assumption that $\neg r$.

[10] At the risk of flogging a dead horse, let me point out that the shifting strict analysis, even if relativised in the obvious way to information states rather than worlds, cannot accommodate this example (as [Fin01] concedes; see his discussion of the "resetting of the modal horizon" by "a rather indirect pragmatic mechanism"). The problem is that the context shift must be assumed to take place for *all* worlds in the belief set; the modal horizons of w (where Big Joe is drunk) and v (where he is not) may gain different worlds, but they both must end up containing counterfactual possibilities where he is drunk. This means that under the shifting strict semantics, factual information *can* have an effect on counterfactuals … but only when that factual information has not already appeared in the antecedent of a previous counterfactual. (Example (8) is predicted unacceptable under the shifting strict analysis, updated to information states, unless you remove "and Big Joe was drunk" from A's first utterance.)

The first element is non-negotiable: in all Sobel-like examples the complex conditional is, ultimately, acceptable. The second is also in accord with normal intuitions, although advocates of the shifting strict analysis would claim that it is a byproduct of the update, rather than an independent fact revealed by growing awareness. It is the third that drives the account for Heim's examples, and the combination of the last two that needs the most justification.

This combination throws light on another feature of Heim's example. There is a temptation to undermine the force of these examples by 'repairing' reverse Sobel sequences. The point is most clearly made with respect not to Sophie but to Lewis's more political example (1-b), expanded to a Heim-style Sobel sequence as follows:

(9) a. (i) If the US threw all its nuclear weapons into the sea tomorrow, there would be war.
(ii) But if the US and all the other nuclear powers threw their nuclear weapons into the sea tomorrow, there would be peace.
b. (i) If the US and all the other nuclear powers threw their nuclear weapons into the sea tomorrow, there would be peace.
(ii) But if the US threw all its nuclear weapons into the sea tomorrow, there would be war.

Discussing this example Gillies writes in [Gil07, p. 332, footnote 5] that there "may be some temptation" to argue that the sequence should instead be read as in (10).

(10) a. (i) If *only* the US threw all its nuclear weapons into the sea tomorrow, there would be war.
(ii) But if the US and all the other nuclear powers threw their nuclear weapons into the sea tomorrow, there would be peace.

Gillies correctly points out that this pair is no evidence against strengthening the antecedent, and so concludes: "We had better resist such temptations."

But the temptation itself is interesting data. (9-a-i), when uttered under the assumption that $\neg r$, expresses exactly the same as (10-a-i). If later conversation makes clear that the utterance *was* made under such an assumption, then a charitable attempt at paraphrase might naturally make that assumption explicit (even if only as a correction: "I thought you meant that if *only*…").

What the first two features above tell us is that in fact these pairs are *not* (in the Heim cases with infelicitous reversals) evidence against strengthening the antecedent for counterfactuals. Such evidence does exist; (4-d) provides it, for example. But what the Heim observations show is that we need to be very careful to enforce *simultaneity* if we are looking for such evidence.

In cases where the Sobel pair is genuinely simultaneously true, the second feature of my summary above fails to hold. In these cases (such as the examples in Section 1.2.1 above), the reversed sequence remains felicitous.

This feature gives a constraint on relative similarity of worlds where Sophie does and doesn't get stuck behind someone tall, and thus indirectly tells us something about the notion of similarity in operation. The ordering semantics is in a sense too flexible: unless we consider only a carefully selected class of possible orderings, we might get stuck with an ordering that considers a world where Sophie at the parade gets stuck behind someone tall less similar to the actual world than one in which she goes to the parade and gets a clear view. Such an ordering would not give us the prediction we want (that the reversed sequence is infelicitous) so we need some way to rule it out. Fortunately, there is a way to pick out the right class of orderings — and even one that has a lot of independent evidence speaking for it. The key is to make use of a *causal* notion of similarity.

3 · *Causal ordering semantics*

Causal semantics for counterfactuals emerges from a different challenge to the similarity accounts of Stalnaker and Lewis: the demand that the primitive notion of 'similarity' be given some content. The most naïve such notion is probably that $w_1 <_w w_2$ (w_1 is more similar to w than w_2 is) just in case w_1 agrees with w on more proposition letters than w_2 does (call this the HAMMING DISTANCE approach). Clearly this will not work if propositions are directly causally related (in w the trigger was pulled and the gun fired; in w_1 and w_2 the trigger was not pulled, but w_1 where the gun still fired is predicted more similar to w than w_2 where it did not). A more subtle problem was pointed out by [Tic76]:

> [C]onsider a man — call him Jones — who is possessed of the following dispositions as regards wearing his hat. Bad weather invariably induces him to wear his hat. Fine weather, on the other hand, affects him neither way: on fine days he puts his hat on or leaves it on the peg, completely at random. Suppose, moreover, that actually the weather is bad, so Jones *is* wearing his hat.

The observation Tichý makes is that "If the weather were fine, Jones would be wearing his hat" seems false in this context, despite (or perhaps *because of*) the lack of any causal dependency between *fine* weather and Jones's habits of hat-wearing.

The basic intuition here is still causal at root: rain causally influences Jones's hat-wearing, so retracting the rain removes the reason for expecting Jones to wear the hat. Veltman gives a more complicated example:

> Suppose that Jones always flips a coin before he opens the curtains to see what the weather is like. Heads means he is going to wear his hat in case the weather is fine, whereas tails means he is not going to wear his hat in that case. Like above, bad weather invariably makes him wear his hat. Now suppose that today heads came up when he flipped the coin, and that it is raining. So again, Jones is wearing his hat.
>
> And again, the question is whether you would accept the sentence '*If the weather had been fine, Jones would have been wearing his hat.*' This time, your answer will be 'yes'. [Velo5, p. 164]

Again the intuition is causal, with an additional wrinkle: rain causes Jones to wear a hat, but in the absence of rain the outcome of the coin-flip will *also* cause him to wear a hat. Veltman models this situation using the notion of the BASIS of a world: a partial propositional valuation minimal such that it picks that world out against the background of live alternatives. If the set of live alternatives does not include every propositional valuation, then some worlds may be picked out without having to give their 'full names' (the complete valuations that identify them). The minimal valuations identifying a world are the bases (there may be several) of that world. In Veltman's scenario above, the actual world is identified by the basis "raining, heads" (against a background of live alternatives constrained by Jones's habits of hat-wearing). Skating briskly over several complications, Veltman's semantics for a counterfactual $\varphi \boxdot\!\!\rightarrow \psi$ are approximated by collecting the worlds satisfying φ whose *bases* differ least (by Hamming distance) from that of the actual world, and testing ψ on that set.

Veltman's system is not truly causal, although it comes very close. Katrin Schulz has improved on it in this respect, in her dissertation [Scho7] and in ongoing work. We need not be too concerned with the details, especially as her proposal is still being refined (for the latest version at the time of writing, see [Sch]). The core idea is to replace Veltman's bases with an explicitly causal notion. The CAUSAL BASIS of a world w is again a minimal propositional valuation (a SITUATION), but one that generates the full valuation at w by the action of default causal rules (such as "rain causes Jones to wear a hat"). In Veltman's system such rules are strict, and act to restrict the universe of possibilities; in Schulz's formulation the world where it is raining and Jones does *not* wear his hat simply has a larger causal basis, recording the fact that some causal expectations have been violated.

Schulz derives an ordering from the relative sizes of causal bases. If b_0, b_1, b_2 are respectively bases for w_0, w_1, w_2, then $w_1 \leq_{w_0} w_2$ iff $b_1 \setminus b_0 \subseteq b_2 \setminus b_0$. That is, if the causal description of w_2 diverges from that of w_0 more than the causal description w_1 does, then w_1 is *causally more similar* to w_0 than w_2 is. Such divergences may be brought about by differences in the facts (that it is not

raining, while in the actual world it is), or by violations of causal expectations ('counting miracles'; for example, that Jones does not wear his hat despite the rain, while in more normal worlds the rain causes him to wear it).

3.1 · *Reasons*

What a causal similarity account brings into clear focus is the importance of *reasons* for the similarity ordering. Our default expectation is that people at parades see the people parading; to not do so *without any reason* is a violation of a causal expectation and would count against the world in question in Schulz's ordering. On the other hand being eaten by a dinosaur would cause Sophie not to see anything; a violation of *this* expectation would also count against the world.

Where things get interesting is, of course, with the worlds where Sophie gets stuck behind someone tall. This also provides a causal reason for her not to see Pedro, but unlike with the dinosaur we want that reason to be 'taken seriously' in the ordering. Schulz's system allows this, as follows. In the actual world (where Sophie is not at the parade) she is not stuck behind anyone tall. But the *reason* she is not stuck behind anyone tall is that she is not at the parade (or in a crowd with tall people). If we counterfactually remove that reason (by moving her to the parade) its causal consequences no longer count for similarity of worlds. Technically, the causal basis of the actual world does not include the fact that Sophie is not stuck behind someone tall (because this is a causal consequence of some other fact in the basis, namely that she is not in a crowd); thus different valuations for this fact do not count to differentiate worlds in the ordering.

(Of course the same does not hold for the dinosaur. The reason Sophie is not eaten by a dinosaur (that they are extinct) applies in both the actual world and worlds where she goes to the parade; changing this fact thus counts as a causal expectation violation and differentiates worlds in the ordering.)

It is very natural in 'counterfactual negotiation' to explicitly bring up reasons ("Sophie's really pushy, she never lets anyone block her view"). I will have more to say about this in Section 4.2, but it should be clear enough that a causal account will behave nicely for such examples. If Sophie's pushiness is causally responsible for her view not being blocked, then a world in which she is blocked violates causal expectations while one in which she is not blocked does not; if she is not pushy (the default assumption), no such violation occurs and the worlds are on an equal footing.

3.2 · *A note of hesitation*

I have resisted giving a full account of Schulz's semantics not just because it is still undergoing revision, but also because I suspect that the fit with awareness is not as tight as I would wish. Her formal implementation relies on

constructing a network explicitly representing the causal influences between atomic propositions. The implementation I have sketched above takes the similarity ordering that such a network produces, and filters that through the assumptions of the agent. While technically quite successful, this approach is conceptually a little suspect. The causal network must be very large and intricate, containing many atomic propositions that the agent does not attend to (it must contain the dinosaur, for instance). It would seem more in the spirit of awareness to filter the network itself, rather than the ordering it produces, through the agent's awareness state. There are however rather formidable technical difficulties standing in the way of this approach; mainly the connection between (possibly complex) assumptions and causal expectations under unawareness is completely obscure, so long as the former are modelled with sets of worlds while the latter are modelled with (something like) formulae.

The notion of closed world reasoning (to which anyway Schulz's technical implementation is closely related) might provide a strategy to overcome these difficulties. Very many of the causal relations we need for examples like Sophie at the parade have the form of a closed-world rule: If you go to a parade then unless something unexpected happens, you see the people parading. (A dinosaur is unexpected at a parade; if it eats you, unless something unexpected happens you won't see anything.) The clause "unless something unexpected happens" is at the heart of the closed-world approach: if we are not told that anything unexpected happened, then assume that it didn't. For causal rules of this form, the relevance of unawareness is clear: the agent need not be aware of all possible exceptional cases in order to reason using the rule.

I have not been able to pursue this hunch further, but the question of the order of explanation seems particularly interesting. It doesn't take much imagination to see a default rule of this kind not as underpinning or generating counterfactual beliefs but rather as an expression or consequence of them. The interaction with assumption becomes particularly interesting; it is certainly no coincidence that the assumption that Sophie is not stuck behind anybody tall has a structure so similar to the formulation as a closed-world rule.

Unfortunately I must leave such speculations, to return to the proposal I am certain enough of to wish to defend. I have claimed as a weakness of the shifting strict semantics that it applies only to counterfactuals (or, on the most charitable reading, only to conditionals). In the next section I will show some non-conditional examples that appear to share features with the Sobel sequence data above, and to which the awareness account can be applied.

4 · *Beyond counterfactuals*

Arguing directly against the shifting strict analysis, Moss [Mos07] gives a number of examples of Sobel-like patterns that do not involve counterfactuals.

Indeed, part of my argument in this chapter is that it may have been a mistake to term the pattern Heim noticed a 'Sobel sequence'. Sobel's observation was that counterfactuals do not support strengthening of the antecedent; if we accept the ordering semantics, then that fact has nothing to do with awareness. Heim's observation, I would say, is that truth-value judgements of counterfactuals can change under conditions of changing awareness; that fact has very little inherently to do with counterfactuals!

So I need to show that my account can be extended to the non-counterfactual cases without difficulties. What are these cases?

Perhaps most famously there is [Lew79, pp. 354–355]:

> Suppose I am talking with some elected official about the ways he might deal with an embarrassment. So far, we have been ignoring those possibilities that would be political suicide for him. He says: "You see, I must either destroy the evidence or else claim that I did it to stop Communism. What else can I do?" I rudely reply: "There is one other possibility — you can put the public interest first for once!" That would be false if the boundary between relevant and ignored possibilities remained stationary. But it is not false in its context, for hitherto ignored possibilities come into consideration and make it true. And the boundary, once shifted outward, stays shifted. If he protests "I can't do that", he is mistaken.

Here, at least, I need do no extra work: modals of possibility and necessity *must* be sensitive to what options are being attended to, as I have already argued for *might* in the previous chapter. But there are a number of non-modal examples which are a little trickier.

[Wilo8] proposes a shifting strict analysis for *indicative* (rather than counterfactual) conditionals, for cases like the following (I adapt the example a little, innocently I hope):

(11) a. (i) If Oswald didn't kill Kennedy that day in Dallas, somebody else did.
(ii) But if the KGB kidnapped Kennedy and his death was faked, nobody killed him.

b. (i) If the KGB kidnapped Kennedy and his death was faked, nobody killed him that day in Dallas.
(ii) ?But if Oswald didn't kill Kennedy, somebody else did.

Moss points out the analogy with a famous example:[11]

[11] Moss admits that "our familiarity with zebra examples can create unwanted noise in our judgements about them." She suggests as a fresh alternative the exchange given in (6).

(12) a. (i) That animal [a zebra] was born with stripes.
(ii) But cleverly disguised mules [with stripes painted on] are not born with stripes.
b. (i) Cleverly disguised mules are not born with stripes.
(ii) ?But that animal was born with stripes.

The original point of these examples was that the possibility being introduced apparently undercuts the ability of the speaker in (12-b-ii) to *know* what is here given as his utterance. That same property seems to make the utterance infelicitous, at least if it is imagined as not stressed in any way (the stressed example belongs with (13) below).

A devotee of update semantics should feel a little uncomfortable with these examples. On the face of it, (11-b-ii) and (12-b-ii) are nothing but bald assertions; the hearer's state should be updated with their information content, and this will cause no problems unless the hearer holds an explicitly contradictory belief. An utterance that would take the hearer to the absurd state is pragmatically ruled out, but that is certainly not the problem here... so what is going on?

4.1 · *Speaker expertise*

The answer is that we are also used to the speaker knowing what they are talking about. The notion of 'an update with the information content of an utterance φ' assumes that φ contains *information*: that what it says is true, or describes the world truly. The very construction of these examples leads us to believe that the utterer of (11-b-ii) or (12-b-ii) is not in a position to know that her utterance is true.

Moss describes this as a norm of assertion: speakers should not make assertions that are incompatible with any salient possibilities that they are not in a position to rule out. She also points out (footnote 10, pg. 11) that "One might aim to derive this principle from others, e.g. from the knowledge norm of assertion and the principle that a speaker knows a proposition only if she can rule out salient possibilities incompatible with that proposition." Awareness-relative epistemology will have to remain a project for future work, however some bounds can be fairly confidently assumed. While I hesitate to say how damaging unawareness may be to an agent's *actual* knowledge, she certainly does not *believe that she knows* that φ (say) if she entertains and holds possible some contingency incompatible with φ. And while I hesitate to say whether knowledge should be a norm of assertion, I'm much more confidant that *believing to know* should be.[12] The awareness-relative reading of all these

[12] All this hesitation is tiresome. But the unawareness perspective seems to tend inexorably towards a particular kind of relativism: *According to A*, B knows that φ. The observer/judge A is needed to set the standards of awareness; otherwise B either knows too much (if her own standards are also normative; "the epistemic efficacy of stupidity", as Catherine Elgin puts it [Elg88]) or too little

examples is that the second speaker of the reverse sequence does not believe she knows what she is asserting, which should explain the infelicity.

This effect is heightened, I think, by a subtle property of the linguistic presentation (as forward and reverse pairs of example sentences). That is the temptation to read (11-b-ii) as *identical to* (11-a-i), and (12-b-ii) similarly as identical to (12-a-i). By this I mean not only that they have the same truth conditions, but that we seem to interpret them as if they were prompted by the same epistemic state on the part of the speaker. The state most naturally assumed to prompt (12-a-i) (i.e., the assumption that no clever disguises are in effect) is clearly insufficient to justify (12-b-ii). If the examples were not presented back-to-back, though, we would not be tempted to assume that the epistemic state of the speaker was the same in both cases: if he feels licensed to assert (12-b-ii) then it is because he is sure that there are no clever disguises involved. The examples are of course carefully picked to make this certainty insufficiently justified for a claim of ('strong', 'philosophical') knowledge, but knowledge is not our primary concern here. Compare these variations (essentially read as adversarial conversations, but on that reading completely felicitous, as far as I can tell):

(13) a. (i) If the KGB kidnapped Kennedy and his death was faked, nobody killed him that day in Dallas.
(ii) I've looked through the KGB historical archives. If Oswald didn't kill Kennedy, somebody else did.

b. (i) Cleverly disguised mules are not born with stripes.
(ii) But if you look closely, I'm quite certain that you'll agree with me: that animal was born with stripes.

c. (i) Cars get stolen all the time here in New York City.
(ii) But *my* car is around the corner. I'm naturally lucky.

I include this last example particularly to undercut the idea that a knowledge norm of assertion has any direct relevance for our judgements of felicity. What seems to be at issue is whether (we are satisfied that) the speaker *believes* she has the required knowledge. If she believes this erroneously (as in (13-c-ii)) we may disagree with her statement, but we feel no temptation to censure her as an uncooperative speaker.

To sum up: I agree with Moss that what is at stake is properly addressing the alternative possibilities that have been made salient, and also that this has something to do with norms of assertion. However we part company when

(if the normative standards include awareness of 'everything there is'). I am not epistemologist enough to know what to do with this position, except to wonder where it leaves the notion of knowledge 'in the abstract', without an explicit ascriber. Hence the hesitations. Section 2.1 of Chapter 7 contains some further hesitations on the same subject, as a suggestion for possible further work.

it comes to what the connection actually is. As far as I am concerned it is a condition of rational belief formation that you not discount salient alternatives without reason; the relevant norm of assertion is "Assert only what you believe you know", and the pragmatic infelicity markings above point to a suspicion of irrational belief formation rather than of deception or similar deliberate violation of norms.

At this point we might wonder, can we play the same game with counterfactuals? Can we rehabilitate reverse Sobel sequences just by taking opinionated speakers? Interestingly, the answer seems to be no. The reasons why not have to do with the extra sensitivity of counterfactuals to changes in awareness.

4.2 · *Uncertainty about counterfactuals*

Suppose one is uncertain whether it is really the case that if Sophie had gone to the parade she would have seen Pedro. It seems there are two kinds of uncertainty one could be suffering from, shown in (14). (14-a) is rooted in lack of knowledge about what *is* the case, while (14-b) is rooted in lack of knowledge of what *would have been* the case in the counterfactual scenario.

(14) a. I don't know whether Pedro in fact showed up at the parade; so I don't know whether Sophie would have seen him if she had gone.
b. Sophie is kinda short, and there were lots of tall people at the parade, she might very well have been stuck way back in the crowd and not seen anything; so I don't know whether Sophie would have seen Pedro if she had gone.

The first example is standard informational uncertainty; we might think of the second as METAPHYSICAL UNCERTAINTY, because it is imposed on us by the metaphysically indeterminate nature of counterfactual alternatives. A similar notion arises in considering the indeterminacy of the future. We might want to represent my uncertainty about the outcome of a coin flip differently depending on whether the coin has not yet been flipped or whether it has been flipped but I have not yet looked at it. Stalnaker famously does not believe in metaphysical uncertainty in this sense, since it leads to violations of conditional excluded middle (at the metaphysical level of truth-conditions); instead he would represent the coin-flip with epistemic uncertainty between world-histories (incorporating future events).

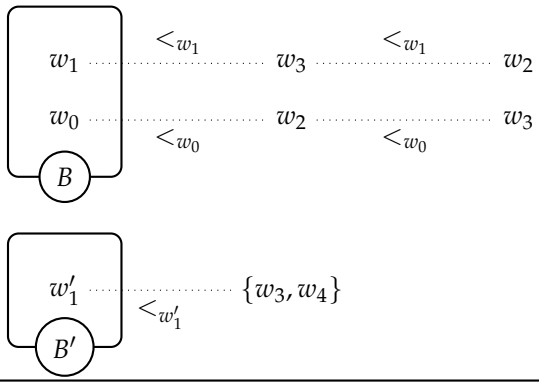
To some extent the decision is an aesthetic one: what individuating conditions for worlds are the most natural in my setting? This is particularly the case for the temporal sequence case, since many branching-time models can be freely inter-translated with models based on world/time pairs or similar constructions. There might be more to the issue for counterfactuals, though. A causal semantics along the lines of Schulz's model, for instance, is *designed* to

generate worlds equidistant from the actual world in cases like Tichý's above: in the absence of reasons for Jones to wear his hat, he may just as well wear it as not wear it. We will see some more examples below where metaphysical uncertainty does some work for us.

Perhaps we can even distinguish between the two notions empirically. To my ears (15-a) is somewhat marked, while (15-b) is fine.

(15) a. ?I don't know whether Pedro was at the parade, so Sophie wouldn't necessarily have seen him if she went.
b. Sophie is kinda short, and there were lots of tall people at the parade, so she wouldn't necessarily have seen Pedro if she had gone.

Here is the picture suggested by the distinction (simplified by omitting several possible valuations). If B represents the agent's state of belief, she would say (14-a), whereas if B' is her state of belief then she would say (15-b).

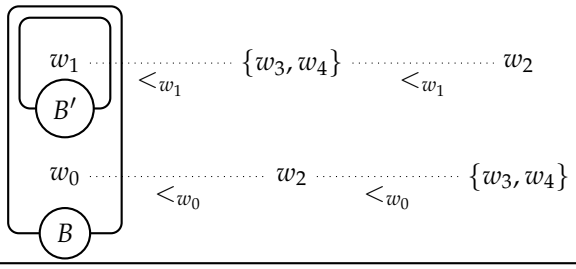


	w_0	w_1	w_1'	w_2	w_3	w_4
Pedro at parade (s)	0	1	1	0	1	1
Sophie at parade (p)	0	0	0	1	1	1
Sophie behind tall people (r)	0	0	0	0	0	1
Sophie sees Pedro (q)	0	0	0	0	1	0

Figure 4.3: Different kinds of counterfactual uncertainty. (Some worlds are shown twice, to make the ordering relations easier to read.) Agents in either of belief states B and B' fail to believe $p \boxdot\!\!\rightarrow q$, however an agent in B can come to believe $p \boxdot\!\!\rightarrow q$ via an update eliminating worlds from her belief set, while an agent in B' cannot.

Although neither information state currently supports $p \Box\!\!\rightarrow q$, there is an important behavioural difference between the two: B can be transformed by an informational update into a state that *does* support the counterfactual, while B' cannot. Suppose for instance that the agent learns that Pedro was at the parade (see Section 2.2). Then w_0 is eliminated, and the resulting state supports $p \Box\!\!\rightarrow q$.

In Figure 4.3 we are to some extent comparing apples with oranges, however: the state rooted at B' assumes that Pedro was at the parade, while the state rooted at B assumes that Sophie would not get stuck behind tall people. If we combine the two into a single state of attention, the result looks somewhat different:



	w_0	w_1	w_2	w_3	w_4
Pedro at parade (s)	0	1	0	1	1
Sophie at parade (p)	0	0	1	1	1
Sophie behind tall people (r)	0	0	0	0	1
Sophie sees Pedro (q)	0	0	0	1	0

Figure 4.4: More complete picture of counterfactual uncertainty.

If the agent assumes $\neg r$ then we get the state labelled B in Figure 4.3; if she assumes s then we get the B' of Figure 4.3. But if she is aware of everything, then even knowing that Pedro was at the parade she remains uncertain whether Sophie would have seen him if she had gone.

If we give anything like a standard update semantics to counterfactuals, then they can be used just like standard assertions to resolve *informational* uncertainty (to eliminate worlds from B) but they cannot be used to resolve *metaphysical* uncertainty. The test semantics we have been using so far does not even allow informational updates, but it does seem reasonable that a counterfactual can carry information ("If Sophie had gone, she would not have seen Pedro" is a roundabout way to tell an agent in state B above that Pedro was not at the parade, on this account). However this possibility is almost completely irrelevant for the vanilla Sobel examples: the 'null context'

of linguistic examples, combined with a reason-based ordering semantics, virtually guarantees that the uncertainty involved will be metaphysical rather than informational.

Definition 4.5: Ordering semantics for attention models (counterfactuals as informational updates). *Let $M = \langle W, \Omega, V, \mathcal{S}^W_{\leq} \rangle$ be a model of reality augmented with an ordering frame, and $\sigma = \langle A, B, \Xi \rangle$ a state.*

$$\langle A, B, \Xi \rangle [\varphi \boxdot\!\!\rightarrow \psi]^M_{\mathrm{b}} =_d \langle A, B^+, \Xi \rangle$$

where

$$B^+ = \{w \in B \,;\, \langle A, C_w, \Xi \rangle \models^M_{\mathrm{b}} \psi\}$$

and for each $w \in B$,

$$C_w = \min_{\leq_w} A \restriction \varphi$$

This is of course nothing but the standard update formulation for assertions: keep the worlds from B that satisfy the formula in question, and throw away the rest.

We have now two update rules for the counterfactual conditional: one a test, and one a substantive (informational) update. One could wonder, do we need them both? I think we do, so long as we make two commitments:

1. counterfactuals can be informative, and
2. the *might* of a *might*-conditional scopes semantically under the conditional operator.

The first principle is uncontroversial; the second is certainly open to debate (an alternative is to stipulate that $\varphi \boxdot\!\!\rightarrow \mathit{might}\, \psi$ be analysed as $\mathit{might}(\varphi \boxdot\!\!\rightarrow \psi)$). If we uphold these two principles, though, we definitely need two kinds of update. The first principle requires (something like) the informational update given in Definition 4.5, but the combination of this definition with the second principle makes *might*-conditionals too strong. In example (7), for example, my utterance would (wrongly) provide the information that Big Joe is drunk.[13]

Assume that we do need both definitions. Then how does an agent, hearing a particular utterance of a counterfactual conditional, choose which one to apply? I think the decision is essentially a pragmatic one, driven by considerations of speaker expertise. Is there a point of epistemic uncertainty at stake, on which the speaker can be taken to be expert? If so, we may apply the

[13] Definition 4.5 is completely out for those who believe in conditional excluded middle: it follows from $\varphi \boxdot\!\!\rightarrow \mathit{might}\, \psi$ that $\varphi \boxdot\!\!\rightarrow \psi$, if no metaphysical uncertainty is possible. The only solution seems to be to scope *might* above the counterfactual; whether this is acceptable (quite apart from compositionality principles) depends on whether you believe a *might*-conditional can ever be informative.

informational update; if not, we had better take the test.[14] There may be more pragmatic reasoning involved, such as whether the counterfactual utterance is a particularly roundabout way of conveying a simple message — I will give an example in a moment, but first now let us see what happens when we try a reverse Sobel sequence in a context where it is clearly epistemic uncertainty at stake (and thus where the informational update would be expected).

(16) a. A: Perhaps Pedro wasn't at the parade. If Sophie had gone and Pedro wasn't there, she wouldn't have seen him.
b. ?B: If she had gone she would have seen him.

My intuition is that (16-b) is somewhat marked, but much more acceptable than (3-b).[15] Part of the markedness is probably due to the fact that speaker B appears to want to convey "Pedro was there" but does to in rather a roundabout way. Moss gives an example which avoids this problem, and also involves informational rather than metaphysical uncertainty:

> Suppose John and Mary are our mutual friends. John was going to ask Mary to marry him, but chickened out at the last moment. I know Mary much better than you do, and you ask me whether Mary would have said yes if John had proposed. I tell you that I swore to Mary that I would never tell anyone that information, which means that strictly speaking, I cannot answer your question. But I say that I will go so far as to tell you two facts:
>
> (18) a. If John had proposed to Mary and she had said yes, he would have been really happy.
> b. But if John had proposed, he would have been really unhappy. [Mos07]

[14] [Vel05] also has dual updates (informational and test) for the counterfactual; for him the choice of which to apply, also pragmatic, rests on whether the laws in play are fully known or not. This is roughly comparable to the question of whether the agent is aware of the causal dependencies at stake or not, although it raises the vexed issue of how an agent *unaware* of some law (or causal dependency) should recognise that unawareness and alter her behaviour accordingly.

[15] It is interesting to note that "But" is completely out here. I have no theory to account for this. As far as I know, nobody has systematically explored in which Sobel-like configurations "but" is permitted, obligatory, or prohibited. For what it's worth, "But" also seems to be out in the dinosaur variant:

(17) a. If Sophie had gone to the New York Mets parade and been eaten by a dinosaur, she would not have seen Pedro Martínez.
b. (? But) If Sophie had gone to the parade, she would have seen Pedro.

Here the feature of the real world to be communicated is Mary's attitude to John, as already conveyed to the speaker in the actual world.

This brings up an interesting point. The best description of that attitude is probably a conditional one: "will not marry him if he asks her". But it is only a very short step to a *counterfactual* attitude: "would not have married him if he had asked her". Indeed, this might very well be what Mary has told the speaker: "For a moment yesterday I thought John was going to propose. I would have said no, though." The feature of the *actual* world that is being conveyed is a *counterfactual* disposition. And if Mary may have this counterfactual disposition to reject proposals, why cannot Sophie have a counterfactual disposition to avoid tall people?

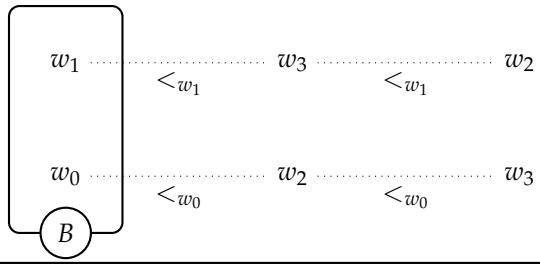
In fact I think she *can*; the two that spring to mind are "Sophie is really pushy and doesn't let anyone get in the way of her view" and "Sophie is really short and always gets a bad view at parades". Either of these, if taken as potential facts in the actual world, would suffice to turn the metaphysical uncertainty about which of w_3 and w_4 takes priority into informational uncertainty (about whether the disposition holds in the actual world). There are two contrasts with the proposal example, however.

The first contrast has already been discussed: Moss's example deliberately makes the speaker an expert on the disposition under discussion, whereas a typical Sobel example undermines the speaker's potential for expertise.

The second contrast is more interesting. This is that both of the dispositions Sophie might hold are 'marked': they need particular names ("pushy", "unusually short") and the most natural awareness model provides global assumptions ("not pushy", "not unusually short") for both of them. In the null context of a Sobel example, it seems we naturally assume (and let our agents also assume) that these dispositions do not obtain. In contrast, the proposal example is explicitly *about* Mary's disposition to accept John's proposal; even if we give the disposition a name that is not being explicitly attended to yet ("love"?) there is no temptation whatsoever to hold assumptions about the valuation that name should get.

Our model is *almost* capable of using this second contrast in order to fully represent the difference between these two situations. The point of leverage is the notion of belief-attention-consistency. The picture corresponding to Figure 4.4 but for Mary's proposal is given overleaf, in Figure 4.5.

Now suppose that the agent is unaware of x, but entertains both w_1 and w_2. By the definition of belief-attention consistency, she may not eliminate either w_0 or w_1 from B! Distinguishing between w_0 and w_1 would violate the clause saying that substantive beliefs may not separate worlds that 'look the same' through the lens of attention: those that make the same propositions of Ξ true. This is a somewhat underhand trick, as can be seen if we recall the reason for



	w_0	w_1	w_2	w_3
John proposes (p)	0	0	1	1
John is happy (q)	0	0	0	1
Mary accepts (r)	0	0	0	1
Mary disposed to accept (x)	0	1	0	1

Figure 4.5: Counterfactual uncertainty for Mary's proposal.

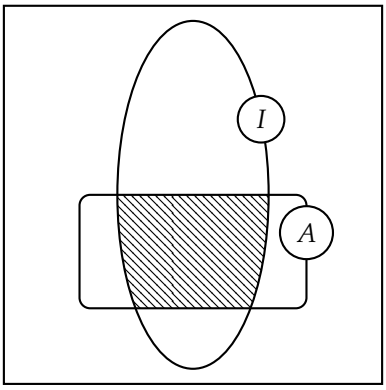
the definition: the agent should be able to *describe* the difference that underlies her substantive belief. In this case she could do so, even given her limited awareness: the formula "$p \boxdot\!\!\rightarrow \neg r$" will do the trick quite nicely. However it is more reasonable if we require the agent to also *justify* her beliefs: the only justification she could give would have to refer somewhere to x.

It is now only a short step (albeit one we will not take *formally*) to distinguishing the two cases. In trying to understand what could justify the speaker in believing "$p \boxdot\!\!\rightarrow \neg r$", our agent *spontaneously becomes aware of* x: the notion of *planning to accept* a proposal is so similar to that of accepting a proposal that this is entirely natural. Now she may form the belief without her state thereby being belief-attention inconsistent. In contrast, in trying to understand why the speaker believes that Sophie would not have been stuck behind anyone tall, our agent fails to imagine anything that could justify the belief; she concludes that something is wrong (either the speaker is making unjustified claims, or she is unaware of something she should be aware of), she rejects the update and instead asks for more details ("What makes you think that?").

Unfortunately this last must remain a Just So story, since we have no mechanism for the spontaneous association of ideas.

Part II

Filtering information with assumptions



In these two chapters our agents have information (in various senses) that is objectively reliable. Their beliefs are formed by interpreting this information through a veil of assumption due to unawareness; as they become aware of new possibilities this veil draws back and their beliefs come closer to a genuine reflection of the information they hold.

Chapter 5

Pragmatics of decision-making

> The question of her decision is one not to be lightly considered, and it is not for me to presume to set myself up as the one person able to answer it. And so I leave it with all of you: Which came out of the opened door — the lady, or the tiger?
>
> Frank Stockton, "The Lady or the Tiger?"

We began this dissertation with a story about Walt's difficult Saturday morning. Late for a job interview and frantically searching for his car keys, he was helped out by Perky Pat:

(1) Did you leave them in the car when you came in drunk last night?

Walt's immediate reaction was to run to the car and check, and by now we have all the tools we need to understand and model his behaviour. Still something is missing: we take Pat's utterance as constituting *advice* for Walt to search the car, but we can only give an ad hoc and intuitive description of why this is so. In a similar fashion, "If you take the bicycle you might get a flat tyre" strikes us intuitively as advice against taking the bicycle, but we would be hard pressed to give an account of why without appealing to the very intuitions we are trying to justify.

A more famous example comes from Grice [Gri67, p. 32]:

(2) a. A: I am out of petrol.
b. B: There is a garage around the corner.

Informally it is easy to see what speaker B intends to convey; Grice derives semi-formally the information that the garage is open via his maxims of cooperative conversation. However a fully formalised account is surprisingly difficult to achieve (in part because of the vagueness of Grice's formulations; as we will see below, trying to formalise reasoning by relevance immediately raises the question of what formal property relevance consists in).

In recent years game theory and decision theory have been fruitfully applied to such problems (see for example the collection [BJR05]), as a formal means of modelling the agent-based nature of discourse (the background context

distinguishing a sentence 'in the abstract' from an utterance made by an agent and to a purpose, which makes pragmatic reasoning possible).

Game theory is perhaps the more popular approach of the two, as it is a particularly good fit for the complex multi-agent epistemic reasoning involved in linguistic coordination, whether conventional [Lew69] or 'online' (e.g., particularised conversational implicature [BR07]). However the very wealth of game-theoretic possibility can be problematic. In coordination games, for instance, standard game-theoretic techniques predict linguistic anti-coordination (in which the speaker makes meaningless noises at random and the hearer ignores him) as a possible 'convention'. Much effort goes into excluding such absurd cases, before we get to the more interesting pragmatic issues. For certain phenomena in pragmatics —notably those involving coordination problems, whether implicit or explicit— we cannot get by without the rich interactional and epistemic space provided by game theory. Before turning to game theory, however, it is interesting to ask whether the questions we are asking can be answered in a model without the rich multi-agent structure that makes game theory both powerful and difficult to work with.[1]

And indeed in our case we can. We will apply the framework of DECISION-THEORETIC PRAGMATICS, which uses the single-agent decision-making model of statistical decision theory, embedded in a multi-agent discourse context that remains mainly notional. The model of decision-making is rich enough to allow complex reasoning about speaker motivations ("Why did she tell me *this*, and not *that*?") of the kind driving the majority of accounts of pragmatic reasoning explicitly based on Grice's work.

It does not, however, explicitly describe unawareness. The representations of previous chapters were all motivated by the intuition that Pat's utterance is intended to make Walt aware of a possibility he is overlooking. The fact that it constitutes advice cannot be separated from this intention, so we will have to find a way to combine decision-theoretic pragmatics with an unawareness model.

As I have argued throughout this dissertation, the best way to achieve such a combination is by the design of 'modular' theories that can be combined without significant alteration. The model I present here falls somewhat short of that ideal; the structures necessary to represent unawareness are significantly more complicated than those usually employed in decision-theoretic pragmatics, and not just in terms of the components strictly necessary for the representation of awareness itself (the atomic formulae attended to, the ordering generating assumptions, and so on). As was the case for counterfactuals (where the ordering structure required by the awareness account gave support for a causal

[1] For a cutting-edge example of both the difficulties and the opportunities provided by rich game-theoretic models, see [Fra09].

analysis of counterfactual similarity) I take this to be no disadvantage for the theory; rather it tells us something more about the right way to think of the simpler structures in common use. Indeed, the parallel between the two cases turns out to be quite close in specific terms, as well as at this level of generality; the counterfactual semantics I gave in Chapter 4 is formally very close to the account of action and agency that underpins the decision-theoretic model.

This chapter grew out of joint work with Michael Franke, which will only see publication after this dissertation is printed and defended [FJo8]. I have revised the formalism substantially, in part to bring it more in line with the systems of the previous chapters and in part because my own views have changed in the interim. I must however record my conceptual debt to Michael; in particular, I owe entirely to him the notion that standard decision theory is riddled through with assumptions about unawareness, which constrain the theory in artificial ways because they remain implicit. It is with this idea that we will begin.

1 · *Decision theory: a theory in need of unawareness*

Bayesian decision theory is a normative theory of decision-making under uncertainty. The decision-maker holds PROBABILISTIC BELIEFS (a probability distribution over possible states of the world) and must choose from a set of possible ACTIONS; a particular state of the world paired with a particular action is associated with a numerical UTILITY expressing relative desirability (thought of as the desirability of performing that action if the world is in that state). The normative theory concerns how these elements should be combined to calculate the best choice of action (I will give more formal definitions in a moment).

A DECISION PROBLEM is a particular instantiation of this pattern, representing a particular choice the decision-maker has to make. For instance, we might model the decisions of an investor choosing which projects and companies to fund, or more prosaically, the choice of how to travel to work in the morning. While awareness is not explicitly a part of a traditional decision problem, it is implicitly absolutely pervasive.

The states of the world that are chosen to represent a particular problem are picked out by attending to some distinctions and ignoring others (the finegrainedness aspect of awareness). The weather is a relevant variable for choosing how to get to work, but completely irrelevant for investment decisions; the modeller must choose which distinctions to attend to and which to ignore. As well as finegrainedness, traditional decision problems embody assumptions: not every conceivable possibility makes it into the states under consideration. Many of these assumptions will be entirely innocent; when choosing how to get to work it is probably safe to assume that there will not be a sudden and devastating flood, or similar madness. However this need not always be the case. An investor may well consider the possibility that each company

she is considering funding might fail, but nonetheless be unprepared for a widespread financial collapse such as the recent credit crisis.

Similar remarks apply to the set of available actions. The set is restricted by something analogous to assumption (there is no point including actions such as "Invest in plastics R&D" if the problem at hand is getting to work on time; less innocently, the agent might be unaware of relevant possible actions, for instance if she does not realise that a new bus line has started operating), and differentiated at a level of finegrainedness that reflects the problem at hand (cycling might count as one action, or be split into the options "take the racing bike" and "take the slow-and-sturdy city bike").

Designing a decision problem, then, is really a kind of attention-modelling task. The modeller must try to choose the right actions and states (drawing attention to the right possibilities) so that the decision problem represents the relevant alternatives while ignoring irrelevant possibilities and distinctions.

Once one takes this perspective, however, an obvious question arises: how are we to represent *changes* in the awareness of a decision-maker, as she comes to recognise new possibilities for action or new ways the world might be? These changes in awareness are an important part of real human decision-making, after all; sometimes the most helpful thing you can do for someone who is struggling with a difficult decision is simply to point out an option they haven't thought of. Because the representation of unawareness in a particular decision problem is only implicit, this kind of change is impossible to model systematically without extending the theory.

In the first part of this chapter I will extend one particular model of decision theory with *explicit* unawareness, so that changes in the awareness state of the agent can lead to evolution of the decision problem she faces. The aim for this section is to enrich standard decision-theoretic representations so that our agents can engage in problem-solving dialogues like those in (1) and (2), and so that changes in the agent's awareness of possibilities systematically carry over into her decision problem.

In Section 4 I will put these enriched models to work on some linguistic issues. Decision-theoretic pragmatics uses the normative theory of decision-making to model pragmatic reasoning about cooperative discourse. By replacing classical decision theory with the awareness-enriched version we can extend the range of decision-theoretic pragmatics without making any changes to the core theory itself. Finally in Section 6 I sum up and mention a particularly interesting loose end.

2 · *Models for decision-making*

We will start with the simple representation of a decision problem which is generally accepted in decision-theoretic pragmatics. (In the wider world of Bayesian or statistical decision theory various alternatives exist, most of which

I will simply ignore for the sake of simplicity.) We need first some basic (and entirely standard) notation for probabilities.

Definition 5.1: Probability basics. *Let S be a countable set. A* probability distribution over *S is a function P from S to the real interval $[0,1]$ such that $\sum_{s\in S} P(s) = 1$. We write $\Delta(S)$ for the set of all such distributions. A distribution on S induces a probability measure on the σ-algebra of subsets of S (also known as* events*), which we also notate with P: if $X \subseteq S$ is any subset of S, then*

$$P(X) =_d \sum_{s\in X} P(s).$$

If P is a probability distribution over S and X, Y are events (subsets of S) then $P(X \mid Y)$ represents the conditional probability *of X given Y, given by*

$$\frac{P(X \cap Y)}{P(Y)}$$

(and of course only well-defined when $P(Y) \neq 0$).

A decision problem formally incorporates the elements given informally above: states of the world and probabilistic beliefs about them, possible actions, and a utility function representing desirability.

Definition 5.2: Decision problem. *A* decision problem *is a structure with four components, $D = \langle S, \mathcal{A}, P, U\rangle$ where*

- *S is a finite set of* states *(not to be confused with possible worlds; states are typically partial objects, recording only a few features relevant for the problem at hand);*
- *P is a probability distribution over S;*
- *$\mathcal{A}$ is a finite[2] set of possible* actions*;[3] and*
- *$U : S \times \mathcal{A} \to \mathbb{R}$ is a utility function: $U(s, a)$ gives the numerical desirability of taking action a if the state of the world is s.*

We will call a state $s \in S$ proper *if $P(s) > 0$.*

[2] Economists are often interested in continuous action spaces representing price offers, production quantities, or other notionally real-valued quantities. The applicability of unawareness to such actions is unclear, since different possible values are not separate concepts in the same way that, say, "cycle to work" and "take the tram" are. Taking a discrete (and finite) action space is also convenient in that it allows me to reuse the awareness machinery developed in the previous chapters without substantial revision.

[3] Note that this is not the assumption set, represented by A in previous chapters. I will adjust the notation for assumptions slightly so as not to mix 'A' and '$\mathcal{A}$', in an attempt —nonetheless probably futile— to avoid confusion.

Given such a problem we can calculate the set of best actions to take by calculating the expected utility of each action (the average utility the action will earn, if the states are distributed according to P). The following definitions are also standard, except for average expected utility (about which more below).

DEFINITION 5.3: Expected utility and related notions. *Let $D = \langle S, \mathcal{A}, P, U \rangle$ be a decision problem, and $a \in \mathcal{A}$ an action. The* EXPECTED UTILITY OF a IN D *is given by*

$$\mathrm{EU}_D(a) =_d \sum_{s \in S} (P(s) \cdot U(s, a)).$$

The AVERAGE EXPECTED UTILITY *of a* set *of actions is the average of their individual expected utilities (equivalent to expected utility under uniform selection of the action): let $X \subseteq \mathcal{A}$, then*

$$\mathrm{AEU}_D(X) =_d \frac{\sum_{a \in X} \mathrm{EU}_D(a)}{|X|}.$$

The BEST ACTIONS *in D are the actions that maximise expected utility:*

$$\mathrm{BA}(D) =_d \{a \in \mathcal{A} \,; \forall a' \in \mathcal{A} : \mathrm{EU}_D(a) \geq \mathrm{EU}_D(a')\}.$$

The VALUE *of D is the average expected utility of its set of best actions:*

$$\mathrm{val}(D) =_d \mathrm{AEU}_D(\mathrm{BA}(D)).$$

The average expected utility of a set of actions demands some comment. It corresponds to a random decision by the agent according to a uniform probability distribution over the set of actions; a biased probability distribution would give a different result, whenever the expected utilities of the actions in the set vary. For calculating the value of a decision problem this is unimportant since all the actions under consideration (the best actions of the decision problem) have the same expected utility (and thus the choice of distribution makes no difference, and the uniform distribution might as well stand in for the others). We will see cases later on, however, where this assumption plays a real role.

Thus far we have decision theory with only implicit unawareness. The tactic I want to apply is to define a richer model (incorporating awareness but also with some additional structure) from which a decision problem can be 'read off'.

2.1 · *Richer decision models*

The eventual aim of this chapter is to be able to model the effect on a problem-solving conversation produced by changes in awareness. Typical conversational moves would be utterances like:

(3) a. Did you think of taking the bicycle?

b. There might be a tram strike.
c. If you take the bicycle you might get a flat tyre.
d. The forecast is for rain; if you bicycle you will get wet.

I will give a model that is rich enough to support the beginnings of a compositional analysis of such updates, although I will not attempt such an analysis in detail. The model is rather complex, so we will construct it in two stages.

In the first stage we define the structures necessary to talk about the state of the world. "If you take the bicycle you might get a flat tyre" talks explicitly about a possible *outcome* of an action, which is nowhere explicitly represented in a decision problem in the format given above. This is no defect for standard decision theory, since the utility is all that is needed for the calculation. However we want our agents to be able to *discuss* outcomes, so they need to be present somewhere in the model. (I will also talk about 'worlds' rather than 'states', for reasons which will become clear in Section 2.2.)

Definition 5.4: Enriched model part 1: states, probabilities, utilities. *An enriched model is a structure* $E = \langle W, \Omega, V, P, \ldots, U \rangle$ *containing the following components:*

W *is a finite set of worlds,* Ω *is as always a set of atomic formulae, and* V *is a valuation function on worlds:* $V : W \times \Omega \to \{0, 1\}$.

P *is a probability distribution over* W*, representing the agent's epistemic uncertainty about the current state of the world.*[4]

$U : W \to \mathbb{R}$ *is a bounded utility function on worlds.*

So far we have enough material for the agent to express her information and uncertainty (via the probability distribution), and her preferences (via the utility function). She is not yet, however, *agentive* in the sense of being able to take action in the world of the model.

Definition 5.5: Enriched model part 2: actions. *An enriched model also contains two more components:* $E = \langle W, \Omega, V, P, \mathcal{A}, O, U \rangle$ *is a fully specified enriched model, where* $\mathcal{A} \subseteq \Omega$ *is a set of special atomic formulae known as* actions*, and* O *is a set of* outcome distributions*, one for each world/action pair. For each world* $w \in W$ *and action* $a \in \mathcal{A}$*, the outcome distribution* $O_a^w \in \Delta(W)$ *is a probability distribution representing the possible results of taking action* a *in the world* w*; if* $O_a^w(w') = 0.3$*, for instance, then taking action* a *in* w *leads to the world* w' *with probability* 0.3*. The only constraint on the distribution is that if* $O_a^w(w') > 0$*, then* $V(w', a) = 1$ *(one guaranteed outcome of taking an action is that the action is taken).*

[4] I will not give any formal representation of temporal structure, however it is important to bear in mind that elements of W may represent either *current* states of the world or states that the world may be brought to through the actions of the agent. The probability distribution represents *current* information; it may be that a world that is assigned zero probability mass is nonetheless reachable via an action.

The outcome distributions reflect the notion of *metaphysical uncertainty* that we saw in Chapter 4. Some actions (such as rolling a die) are inherently indeterminate, and the most we can say about their results is that they will be distributed according to a particular statistical pattern. Of course *epistemic* uncertainty about the results of actions also has a place in the model: if the die may or may not be loaded, then the outcome distributions differ for the world where it is fair and the world where it is not. We will return to this notion below.

Besides metaphysical uncertainty, the main idea of the definition is that actions *change the world*. Trivially, taking the bicycle to work changes the world into one in which the agent took the bicycle to work; typically various other elements of the world will change along with this one (she may get windblown hair and tired legs, and the bicycle will end up locked in the carpark rather than at home in the cellar, for instance). If h is the propositional formula saying that the agent has windblown hair, and $[\![h]\!] \subseteq W$ is the set of worlds satisfying that formula (the event, or proposition), then it is quite reasonable that in her decision problem $P([\![h]\!]) = 0$). This means that her information tells her that *at this moment* she does not have windblown hair; similarly, at this moment she is not bicycling to work (the action b). The conditional probability $P([\![h]\!] \mid [\![b]\!])$ is undefined (since $P([\![b]\!]) = 0$) and anyway does not represent the probability that bicycling leads to windblown hair; it represents the probability that the agent has windblown hair *now*, conditional on her *now* being on a bicycle.

The probability that bicycling to work leads to windblown hair is read off instead by first *changing* the model with action b, then looking at the probability of h in the resulting distribution. The notion was called GENERAL IMAGING in [Gär82],[5] and is defined (in our setting) as follows.

DEFINITION 5.6: Enriched model: results of action. *Take an enriched model $E = \langle W, \Omega, V, P, \mathcal{A}, O, U\rangle$, and any action $a \in \mathcal{A}$. Then $E[a]$, the result of taking action a in E, is a new enriched model:*

$$E[a] =_d \langle W, \Omega, V, P[a], \mathcal{A}, O, U\rangle$$

in which the probability distribution over worlds is given by:

$$P[a](w') =_d \sum_{w \in W} (P(w) \cdot O_a^w(w'))$$

This updated probability represents the chance of *arriving* at w' by performing action a, while the original P represented the chance that the world is *already* w.

[5] Lewis defined imaging in [Lew76], however he was applying it to *Stalnaker*'s counterfactual semantics, with its assumption (validating conditional excluded middle) that each world has only a single closest φ-neighbour. The generalisation given by Gärdenfors allows sets of φ-neighbours, in our case representing metaphysical uncertainty about the outcome of an action.

It is this updated probability we will use in calculating the expected utility of an action.

Seen in this light, the outcome distributions are closely related to a counterfactual similarity relation, especially when the latter is generated by a causal semantics as I have argued in the previous chapter. The counterfactual closest φ-worlds to w incorporate the minimal expected causal consequences if w had been adjusted so that φ was the case; the decision-theoretic b-outcome worlds of w incorporate the minimal expected causal consequences if w is adjusted by *making* b the case. Outcome distributions are simply a probabilistic incarnation of the same idea of causal similarity driving the counterfactual semantics that we have already seen (indeed, imaging was first proposed for a probabilistic treatment of counterfactual beliefs).[6]

There is one significant element left out of the model: temporal structure in the worlds themselves. An element of W should really be a world-time pair, so that our agents can properly discuss knowledge about the past and future; taking an action does not, in fact, 'change the world' in its entirety but only the available *future* course of events, and so on. I will deal with these complications mainly informally; the only temporal dimension explicitly represented in the model is the distinction between the 'now' of epistemic uncertainty and the 'future' of results of actions.

With that proviso, let us see the connection between the enriched model and the decision problem as standardly conceived.

Definition 5.7: Impoverishing an enriched model. *Let E be an enriched decision problem with $E = \langle W, \Omega, V, P, \mathcal{A}, O, U\rangle$. Define the* impoverishment *of E, a standard decision problem $D_E = \langle S_D, \mathcal{A}_D, P_D, U_D\rangle$, as follows:*

- *$S_D = W$ (the states are simply the worlds from the enriched model);*
- *$\mathcal{A}_D = \mathcal{A}$ (the same actions are reused);*
- *$P_D = P$ (the probability distribution representing epistemic uncertainty is also reused);*

[6] Outcome distributions are more direct probabilistic analogues of selection functions: for each world w and formula φ, $f(\varphi, w)$ gives directly the closest world(s) to w where φ holds. (Stalnaker introduced the notion on the assumption that a single world was closest, validating the conditional excluded middle; Lewis generalised it to allow a set of equidistant closest φ-worlds, which is the formulation we take here.) Every similarity ordering can be represented by a family of selection functions, but the converse does not hold. An ordering carries additional information about similarity, such as that it is transitive, which must be expressed as a constraint on sets of selection functions. I assume implicitly, for simplicity in discussion, that actions are mutually incompatible, so that there is never any reason to compare the relative similarity of the closest a-world and the closest b-world to w (where a and b are actions). Allowing mutually compatible actions would simply require putting the same constraints (expressed probabilistically) on combinations of outcome distributions that similarity orderings impose on selection functions.

- *U_D is given by, for each $w \in S_D$ and $a \in \mathcal{A}_D$:*

$$U_D(w, a) =_d \sum_{w' \in W} (O_a^w(w') \cdot U(w')).$$

The impoverished decision problem is clearly very closely related to the enriched model it derives from. The only real difference between the two representations is that the impoverished model hides the detailed structure of (expected) outcomes behind the numerical calculation of utility. However, this kind of impoverishment, with all its fine detail, is not generally what we will want. An enriched model contains full possible worlds rather than partial states, so it will typically be much more fine-grained than our agent's mental state. We need a way to turn the total structures of the enriched model into partial structures, keeping only the distinctions (and actions) that the agent is aware of.

2.2 · *Impoverishment via unawareness*

Just as for the dynamic model of Part I, the strategy will be to treat the enriched model as a 'model of reality', and restrict the agent's view of this model according to her state of (un)awareness. Unlike in Part I, however, the emphasis here is on partiality rather than assumption. Typically the states of the impoverished decision problem will aggregate a large number of worlds, erasing the distinctions between them in the awareness-limited view of the agent. And just as I argued in Chapter 1 for unawareness of objects, the 'gap model' (with unawareness corresponding to an 'assumption of absence') is particularly appropriate for the actions of a decision problem. If the agent is unaware of the possible action "Take bus 405" (whether because she has not thought of buses at all, or because she does not know of that particular line), that action should simply be absent from the impoverished model representing her view of the decision problem she faces.

The two modes of unawareness (assumption and finegrainedness) correspond to two kinds of filters that we will apply to get from an enriched model to an impoverished decision problem under unawareness: assumptions provide RESTRICTIONS, and finegrainedness provides AGGREGATION.

DEFINITION 5.8: Awareness state. *Fix an enriched model $E = \langle W, \Omega, V, P, \mathcal{A}, O, U \rangle$. Similar to the definition of Chapter 2, an* AWARENESS STATE *is a pair $\sigma = \langle W_\sigma, \Xi \rangle$ where*

- *$W_\sigma \subseteq W$ represents the assumptions (relabelled so as to avoid both 'A' and '$\mathcal{A}$' as components of the similar structures);*
- *$\Xi \subseteq \Omega$ is as before the set of proposition letters the agent attends to.*

Recall that $\mathcal{A}$, the set of actions, is nothing more than a specially-treated subset of Ω. Likewise, we can single out the actions the agent is aware of: $\mathcal{A}_\sigma =_d \mathcal{A} \cap \Xi$.

Unlike the awareness states of the previous chapters, we don't need to include beliefs. These are encoded in the prior probability distribution over the enriched model as a whole, and we will derive probabilistic beliefs under assumptions simply by conditionalising on the assumption set W_σ. The idea is that the distribution over the enriched model records the probabilistic beliefs the agent *would* have, if she paid proper attention to all possibilities — a sort of limit case of total attention. To make this precise, we need to see how to read off an impoverished decision problem via an awareness state.

Definition 5.9: Impoverishment via unawareness. *Let E be an enriched model with $E = \langle W, \Omega, V, P, \mathcal{A}, O, U\rangle$ and $\sigma = \langle W_\sigma, \Xi\rangle$ an awareness state. Then the impoverishment of E under σ is a decision problem*

$$D^\sigma_E = \langle S_D, \mathcal{A}_D, P_D, U_D\rangle.$$

We need a subsidiary notion to define the states. Let $w_1, w_2 \in W$ be two worlds. Then w_1 and w_2 are equivalent under Ξ*, written $w_1 \equiv_\Xi w_2$, if for each $p \in \Xi$, $V(w_1, p) = V(w_2, p)$. Now the components are given by:*

- *$S_D = W_\sigma/\equiv_\Xi$; that is, the states are the equivalence classes under Ξ of the worlds the agent entertains;*
- *$\mathcal{A}_D = \mathcal{A}_\sigma$ (the actions of the decision problem are just those the agent is aware of);*
- *P_D is given for each $s \in S_D$ by*

$$P_D(s) = \sum_{w \in s} P(w \mid W_\sigma) = P(s \mid W_\sigma)$$

  (that is, the probability distribution treats states as events in the enriched model, conditionalised on the agent's assumptions); and

- *U_D is given for each $s \in S_D$ and $a \in \mathcal{A}_D$ by*

$$U_D(s, a) = \sum_{w \in s} \sum_{w' \in W_\sigma} (P(w \mid s) \cdot O^w_a(w' \mid W_\sigma) \cdot U(w')).$$

The utility calculation contains nothing unexpected. To find the expected utility of action a in state s we need to take expectations over worlds in s (epistemic uncertainty) but also over outcomes of the action in those worlds (metaphysical uncertainty), before measuring the utility in those outcomes. The most important point is the conditionalisation: the agent cannot see outside W_σ, so the outcome probabilities need to be scaled accordingly (an agent unaware of the double-zero in American roulette would assign the outcome probability $\frac{1}{37}$ to each of the other numbers zero through 36 on the wheel).

The impoverishment transformation encodes the two modes of unawareness: the decision problem includes only worlds from the set W_σ (assumption), and the states are distinguished only according to propositions in Ξ (fine-grainedness). Whenever a state contains worlds differing on the value of some proposition letter p, we can see the state as a *partial* object unspecified for the value of p. However if the agent later becomes aware of p, the partial object that is the state splits into two more completely specified objects, differing in their valuation for p.

The resulting decision problem also has the intuitive property that the agent can describe the differences between the different states, using only her language $\mathscr{L}^\Xi$. However, just as for awareness states in the previous chapters, it may be malformed in more subtle ways.

Definition 5.10: Awareness-consistency. *Let $E = \langle W, \Omega, V, P, \mathcal{A}, O, U\rangle$ be an enriched model and $\sigma = \langle W_\sigma, \Xi\rangle$ an awareness state.*

Just as for our original models, a state is awareness-consistent *if every valuation of Ξ that occurs in the model also occurs within the assumption set: for all $\varphi \in \mathscr{L}^\Xi$,*

$$\textit{if for some } w \in W,\ w \models \varphi, \textit{ then for some } w' \in W_\sigma,\ w' \models \varphi.$$

The state is decision-awareness-consistent *if*

$$\exists w \in W_\sigma : P(w) > 0$$

(the assumptions permit probabilistic beliefs; equivalently, $P(W_\sigma) > 0$), and

$$\forall w \in W_\sigma\, \forall a \in \mathcal{A}_\sigma\, O_a^w(W_\sigma) > 0.$$

(results of actions are always probabilistically defined).

Just as for the simpler models of the previous chapters, we need to define an ordering $\preceq$ on W that provides the assumptions. Not every ordering will produce decision-awareness-consistent impoverished models, because of the two conditions. A natural way to force the assumptions to permit probabilistic beliefs is to require that some elements of the lowest level of the ordering carry probability mass. (Recall that regardless of the awareness state of the agent, the worlds in the lowest level of $\preceq$ are always included in her assumptions.) It is natural that *some* of these worlds are held possible because of the intuitive correspondence between probabilistic belief and assumption: the lowest level of $\preceq$ represents the way the agent assumes the world is before she gives the matter any thought. If her probabilistic beliefs as encoded 'in the limit' of the enriched model should not give these worlds probability mass, this would be at least evidence that her assumption-formation machinery is operating very

badly.[7] On the other hand it is equally natural that some of these lowest-level worlds be held (probabilistically) impossible (and thus a requirement that all the worlds carry probability mass would be too strong): they may represent *future* situations that are so salient the agent automatically entertains them, despite only holding them possible as the result of actions not yet taken.

More problematic is the requirement that action results are everywhere defined. It might seem that awareness-consistency should guarantee this, since if the agent attends to an action a, it provides worlds in which the action a is taken (that is, where $V(w, a) = 1$). The problem is that these need not necessarily be the outcomes of worlds the agent holds possible. For a trivial example, if all the worlds in W_σ that satisfy a are (notionally) in the past of w, and the agent holds w possible, then from w the action a has no outcome in W_σ (since actions cannot 'time travel' into the past). I will not attempt to solve this problem with a principled restriction on the structure of $\preceq$ (if such is possible, the expression would seem to be rather complex); instead I will simply require that all impoverished models satisfy the constraints. This will be simple enough to verify by hand, for the examples I am interested in.

3 · *Updates*

The system is to some extent a direct descendant of the models of the first part of the dissertation, so it could in principle be applied to the same kinds of problems. The extra structure of decision theory, however, lets us model advice-giving that talks explicitly about the opportunities for action available to the agent, as in (3).

I will be vague about the specifics of the update mechanism, since it follows the pattern of the simpler system rather closely. For attention updates we replace W_σ in the awareness state with an ordering $\preceq$ on W, and generate W_σ from $\preceq$ and Ξ as in the non-probabilistic model. Belief updates also operate analogously, with non-zero probability corresponding to membership in the belief set B (so that "*might* φ" holds if some world supporting φ gets non-zero probability, and the update with a propositional formula φ sets the probability of each w_t not supporting φ to zero).

The interesting complication is the action update "If you take the bicycle, you might get a flat tyre." For this we introduce the connective $\Box\!\!\rightarrow$, now representing not a counterfactual but an action update (it must be syntactically restricted to take only an action in the antecedent, not a general formula). The reuse of notation is entirely justified, however, since the semantics is exactly analogous to that given in the previous chapter (Definition 4.5, the informational

[7] This constraint only makes sense when the prior encoded in E represents the 'start of a conversation'. As the agent accumulates information she will naturally be able to rule out worlds in the lowest level of the association ordering, which we will implement with her updating the prior by conditionalisation.

update version).

That is, an enriched model E supports an 'action conditional' $a \boxrightarrow \varphi$ if $E[a]$ (probabilistically) supports φ. Updating with $a \boxrightarrow \varphi$ is only possible if φ is propositional (does not contain *might*) since the update works by eliminating epistemic uncertainty (removing worlds from the set of current epistemic alternatives).[8] In particular, "If you roll that die you will get a four" is only acceptable if some epistemic alternative contains an unfair die; it cannot be interpreted as pruning outcome distributions (resolving metaphysical uncertainty) but only at the epistemic level.

This is perhaps the simplest probabilistic update system that can be imagined (modulo the awareness dimension and the complexity of the model itself!): propositional updates correspond simply to conditionalisation. In particular, although the model can represent subtleties such as "φ is more likely than ψ", neither we nor the agents can talk about these systematically (since the object and meta languages are none other than those defined in Chapters 2 and 3).

The awareness updates, however, bring to the fore one final subtlety, which I mention for completeness even though it will play no part in the rest of the chapter: becoming aware of one proposition letter may trigger awareness of another, via a process of ASSOCIATION.

3.1 · *Associations*

Associations were introduced into the formal system of [FJ07] to solve a technical problem. In that paper we distinguished ontologically between actions and proposition letters; becoming aware of an action (such as taking the bicycle) then needed to be associated with becoming aware of at least one potential outcome for that action (having taken the bicycle, as a proposition). This is no longer strictly necessary in the current model, since the action does double duty as a proposition, however it is certainly natural, particularly in cases where an action has two 'stereotypical' or 'expected' outcomes. Suggesting that the agent flip a coin, for instance, can be reasonably expected to make them aware not just of the action but of the outcomes 'heads' and 'tails'. In fact none of the examples I will consider below have this property, so I will refrain from adding a representation to the formalisation. The notion should probably be added to the simpler models of the previous chapters; a simple way to do so would be to give each proposition letter a set of associations (also proposition letters, trivially including the letter itself), and define the awareness update with φ not based on the proposition letters occurring in φ but on the union of their

[8] This is an unfortunate consequence of the 'double uncertainty' of these models: a single *might* can represent either epistemic or metaphysical uncertainty, or even both (recall "If I had taken the fight, I might have won"). The informational update rule that makes propositional updates work gives the wrong results for might conditionals, while the test rule that gives might conditionals the right semantics fails for updates. See the discussion following Definition 4.5 of Chapter 4.

associations.

3.2 · Examples

The most important result of this proposal is that we can accommodate the effect of *uninformative* utterances, in the typical awareness mode. I want to first describe the effect of simply becoming aware of a possibility, without this being induced by hearing a linguistic utterance. The pragmatic effects of uninformative utterances are so important that they tend to overwhelm any intuitions about what other effects they might have. I will try to avoid this effect by framing each example as "Oh! It's possible that…" or similar; the idea is that the agent has spontaneously become aware of a new possibility, without that possibility being suggested to her by someone else. In Section 4.3 below I show the kinds of pragmatic reasoning this can give rise to.

3.2.1 · Oh! I could take the bicycle!

This awareness update obviously adds an action to $\mathcal{A}$ (say, b for "bicycle"), but it must also update the worlds in W_σ accordingly. Becoming aware of b will very likely add some worlds to W_σ, and thus change the set of states, but it need not alter the *proper* states (those that carry probability mass) at all. Remember that these represent what is the case *now*, while taking the bicycle is an action that (typically) occurs in the future.[9] Typically, then, the worlds added by the awareness update (and the states they aggregate into) will carry zero probability mass in E but positive mass in $E[b]$.

The proper states, then, need not change at all, but quite possibly some new worlds will need to be added to W_σ, to ensure that every world currently held possible has an outcome (carrying non-zero probability mass after the action update) for b. In addition, the action itself becomes available and thus the linkage (via outcome functions) between current states and possible futures becomes richer.

Looking at the impoverished decision problems before and after, the states and probabilities probably will not change; the set of actions grows, and the utilities of the other actions besides b are unlikely to change (only if the states do, in fact), but of course the utility of action b becomes defined at each state.

3.2.2 · Oh! It's possible that there is a tram strike!

This update has different effects depending on what the agent assumed before she was aware of the possibility of a tram strike. We consider two cases:

[9] I have to hedge with "typically" and so on because it is possible, in principle, to model an agent who becomes aware that she can take a bicycle to work, and simultaneously realises that she is not sure if that is what she is *already doing*. A famous problem in game theory, the 'Drunk driver paradox', concerns a similar scenario (without the awareness dimension) so the notion, though patently absurd, has some application. We will not do any work with such marginal examples, but it is reassuring to know that the system can represent them if pushed to it.

either she was assuming that there would *not* be a tram strike (a reasonable assumption; the same account holds, *mutatis mutandis*, for the unreasonable assumption that there *would* be a strike), or she held no such assumption and the finegrainedness effect of unawareness simply left her states unspecified for strikes.

In the first case, the new states she gains will be entirely disjoint from the old ones. The relative probabilities of the existing states will be left unchanged (so if the chance of rain was $\frac{2}{3}$ on the assumption that there is no strike, the chance of rain conditional on there being no strike is still $\frac{2}{3}$). Their absolute probabilities, however, will change as the new states take up probability mass, affecting calculations of best actions and so on.

If the agent did not assume there would be no strike, matters are slightly more complicated. Some or all of her states will have to *split* into halves holding striking and non-striking worlds (and there may still be entirely new states included, for the same reasons as in the simpler case). For a case like a tram strike, splitting states will reveal widely varying utilities across worlds that used to be considered part of one state (taking the tram is of course strongly dispreferred if it doesn't actually go where you're going…). My feeling is that for such stark utility differences an assumption is a more realistic modelling choice, because the finegrainedness solution involves averaging over conceptually highly distinct outcomes which the agent nonetheless does not distinguish. On the other hand it is easy to imagine an agent who has not realised that Tram 12 takes two different routes, one of which is longer than the other; her assessment of the utility of "take Tram 12 at 13.10" is naturally seen as an average over her uncertainty about which route that particular service takes, with the averaging kept invisible to her by her unawareness of the distinction.

3.2.3 · *Oh! If I take the bicycle I might get a flat tyre!*

The final example of an awareness update concerns a possibility that can only be realised if the agent takes a particular action.[10] Mechanically speaking the result is very similar to "There might be a tram strike": the newly raised possibility adds more worlds, with concomitant changes to the set of states and the possible outcomes. As for adding an action, however, we would typically expect the distribution on (proper) states to remain largely unchanged: the possibility being raised belongs in a future that is only accessible via 'changing the world' with the action of taking the bicycle.

So far these cases are relatively straight-forward extensions of update semantics with unawareness into decision theory. We are not yet really modelling problem-

[10] In fact the awareness update will function just as effectively if there is no connection between antecedent and consequent, as in the clearly infelicitous "If you take the bicycle there might be a tram strike". This issue should probably be resolved by pragmatics but, in contrast to pragmatic questions of sufficient probability and utility, the proposal I give in Section 4 does not achieve this.

solving discourse, as in (3), though, because these updates completely ignore the pragmatic force of the utterances. In the next section we take the necessary further step: the assumed background of a decision problem can provide a representation of *context of utterance*, allowing our agents to perform pragmatic reasoning.

4 · *Decision-theoretic pragmatics*

Decision-theoretic pragmatics emerges from the observation that conversation is very often 'to the point', where the point in question is some non-linguistic aim; this is particularly the case in information-gathering scenarios such as Grice's famous example given in (2).

In cases like these we might model the first speaker's situation as a decision problem ("How to get petrol") with utilities corresponding to eventual success; her conversational aim is to increase the value of this decision problem, which is nothing but a formal way of saying she aims to maximise her chances of getting her car refueled.

Grice's seminal notion of COOPERATIVITY can be modelled in such a setting by assuming that the utilities of A's decision problem are shared by all conversational participants. However other notions at play in the Gricean maxims have proven somewhat more resistant to formal explication. The key target of decision-theoretic pragmatics is a formal representation of the Gricean notion of RELEVANCE. [Mer99] gave an explication in terms of resolving a background question, while Prashant Parikh and Robert van Rooij[11] have described fully decision-theoretic models where comparative relevance corresponds to effectiveness in resolving a decision problem [Par01; Roo03a; Roo03b].

4.1 · *A measure of relevance*

Van Rooij argues for a measure of relevance given by the VALUE OF SAMPLE INFORMATION (VSI), a notion from statistical decision theory (see e.g. [RS61]). We will use this measure; I give now an argument that derives it.

Suppose the agent is facing a decision problem D, and receives the information that φ. We will set awareness updates aside for the moment, so the new decision problem can be written $D[\varphi]_b$ (D transformed by a *pure belief* update with φ, implemented by conditionalising on the support of φ).

[11] The observant reader will note that the references are to papers by "Robert van Rooy". The spelling variation is a quirk of Dutch orthography. The modern Dutch alphabet contains a letter known as the 'long y', variously written *y*, *ÿ* or *ij* (the distinction between the latter two vanishes in handwritten cursive). Confusingly to the English-speaking reader, IJmuiden (a port close to Amsterdam) is so spelled (and capitalised!) but may be found alphabetised, in a Dutch index, just before Y. (Or it may not. Usage still varies a lot within the Netherlands, with even subject matter having some influence — telephone directories list Cruijff next to Cruyff, while modern dictionaries would split them, filing IJmuiden under I.) Such pedantry aside, Robert van Rooy and Robert van Rooij are the same person, who nowadays prefers the latter spelling.

Intuitively, the value of the information that φ should be something like the difference between the expected values of her decision problem before and after learning that φ. Suppose we implement this naïvely: the value of φ is simply $\mathrm{val}(D[\varphi]_!) - \mathrm{val}(D)$ (recall that $\mathrm{val}(D)$ is the expected value of any best action in D, in other words, the utility the agent can expect if she makes the best possible choice of action). But now 'unwelcome information' has negative value! If there is only a 0.01% chance that I have cancer, the expected utility of taking no treatment can be quite high ($\mathrm{val}(D) = 10$, to name an entirely arbitrary figure); but if I learn that I *do* have cancer, my best action may be to take an expensive and painful chemotherapy treatment ($\mathrm{val}(D[\varphi]_!) = 2$, incorporating both the unpleasantness of the treatment and the chance that it nonetheless is unsuccessful). Because the news that I have cancer confirms that the state of the world is an inherently bad one, according to this measure the information has negative value; I would rather *not learn* that I have cancer, even though remaining in ignorance will lead me to take an action (not seeking treatment) leading to an unpleasant death ($\mathrm{BA}(D) = \{\text{no treatment}\}$, but $\mathrm{EU}_{D[\varphi]_!}(\text{no treatment}) = 0$).[12]

The value of sample information instead recognises that if φ is in fact the case, however unwelcome the *fact* may be, the *information* (that φ) is never of negative value. On the standing assumption that the information (that φ) is true, it establishes the fact that φ; the values to be compared are the expected utilities of acting with or without the information that φ, but on the standing assumption that (the fact that) φ holds. The definition is as follows:

Definition 5.11: Value of sample information (VSI).

$$\mathrm{VSI}_D(\varphi) =_d \mathrm{EU}_{D[\varphi]_!}(\mathrm{BA}(D)) - \mathrm{EU}_{D[\varphi]_!}(\mathrm{BA}(D[\varphi]_!)).$$

That is, we compare two sets of actions: those that were considered best in the original decision problem ($\mathrm{BA}(D)$), and those that are considered best after learning that φ ($\mathrm{BA}(D[\varphi]_!)$). But we take the expected utilities of all of these actions with respect to the *updated* decision problem $D[\varphi]_!$ (note the subscripts on the expected utility calculations). VSI represents the value of *receiving the information* that φ, whereas the naïve measure above represented the value of *bringing it about* that φ; bringing it about that I have cancer has negative value, but if I *do* have cancer then learning that fact has positive value.

[12] This unexpected consequence seems to have been overlooked in early work in the field (see for example [Par92] for an implicit use of the naïve notion in a game-theoretic setting). Probably this is because of an early focus on game-theoretic models of pure coordination, in which it is the joint behaviour of the players that largely determines the utility outcome, rather than the state of the world (there is usually no such thing as inherently bad news in such a model).

4.1.1 · *Some properties of interest*

The first notable property of VSI has already been alluded to: it is non-negative. Information may be irrelevant (when $\mathrm{VSI}_D(\varphi) = 0$) but never actively harmful. I will mention one apparent counterexample, in order to dismiss it (and a similar class of objections). Jim has a wife Jane and a mistress Joan, and neither of the two women know of the existence of the other. One day by coincidence both Jane and Joan happen to visit the same café. Now the information that Jane is in the café would lead Jim to go in himself, leading to a very unpleasant scene; surely this information should have negative value?

The argument relies on a refusal to take probabilistic expectations seriously enough. If Jim prefers to visit the café on the strength of the information that Jane is there (and taking into account the possibility that Joan might be as well), the chance that both women are there at the same time must be very small. It is that chance that the calculation makes reference to. The fact that the actual world happens to inhabit that low-probability region is simply a case of bad luck for Jim; over a large enough sample of alternative worlds this bad luck would 'average out' to approximate the expectation calculation.[13]

It is easy to see that in such cases of 'dangerous information', more specific information that 'resolves the danger' will always have a higher value. In this case, both "Jane and Joan are in the café" and "Jane is in the café and Joan is not" have a higher value than "Jane is in the café", in the respective worlds where they are true; the former because it leads to a more appropriate action, and the latter because it eliminates the small possibility that the action taken —entering the café— will lead to disaster. The situation is slightly different when φ can be an uninformative utterance leading to an awareness update; I will return to the point in Section 6.1.

Another property of VSI is that it can be non-zero only if the set of best actions changes between D and $D[\varphi]_{\mathrm{b}}$. According to Van Rooij, "it doesn't seem unnatural to say that a cooperative participant of the dialogue makes a *relevant* assertion in case he influences the action you are going to perform" [Roo03b, 735, orig. ital.]. On the other hand, the measure seems to be too strong, in that information can be intuitively relevant even if it does no more than confirm the optimality of an action taken under uncertainty. [Roo03b] acknowledges this point, but cites the naïve measure given above (also used by [Par01]) as a solution, which we have seen cannot be suitable. A more promising approach might acknowledge the uncertainty we have, as decision-makers, *about the values of our probabilistic uncertainty*. Given such higher-order uncertainty, small differences in expected utility (between two actions under consideration) are a risky basis for decision-making, while large differences are more robust and

[13] See however Section 6.1 for a suggestion that this argument does not carry over to the case of awareness.

reliable; I will not, however, introduce higher-order uncertainty into the model.

One final note on formalism, before we turn to the applications: it is because the best actions of D are evaluated in $D[\varphi]_b$ that we need a definition for the average expected utility of a *set* of actions. By definition, the best actions of D all have the same expected utility in D; however they may differ in effectiveness in the updated decision problem. The definition we have given corresponds intuitively to an agent who chooses randomly (with uniform distribution) between actions with the same expected utility. In reality such choices might also be influenced by other factors, in particular by considerations of risk (in the same way that in reality information that simply reinforces a decision may still be intuitively relevant). Again, I will not model this possibility; intuitively, however, the effect would seem to be to slightly raise the value of 'extreme' information favouring the risky alternatives.

4.2 · *VSI for pragmatic reasoning*

The use of VSI in reasoning is based on a PRESUMPTION OF RELEVANCE: the hearer assumes the speaker has followed the Gricean maxim of relevance, and tries to find an interpretation of the utterance that makes this the case. (It is thus unsuitable for deriving implicatures based on *flouting* the maxim, in which the semantic meaning is genuinely irrelevant.)

Consider the Gricean example "There is a garage around the corner". Van Rooij writes [Roo03a, p. 1175] "Because B's reaction can only resolve these issues [i.e., the decision problem "Where can I get petrol?"] when the garage is open, A understands that this is conversationally implied by B; otherwise the relevance [under VSI] of his assertion would be 0, i.e. his information would be pointless." This argument is parallel to Grice's own (although I will argue below that this example is better thought of as an *awareness* update) but the decision-theoretic framework allows some more subtle effects that are otherwise difficult to capture informally without ad hoc argumentation. Two examples (taken from two papers of Van Rooij) will suffice to give the flavour, although I will give only the most cursory presentation; for more examples and a more nuanced consideration of the possible complications see the respective publications.

4.2.1 · *Generalised quantity implicatures [Roo03a]*

A QUANTITY IMPLICATURE arises from Grice's (first sub)maxim of quantity, "Make your contribution as informative as required for the current purposes of the exchange." Early work on these implicatures focused on so-called IMPLICATIONAL SCALES such as ⟨some, most, all⟩ and ⟨or, and⟩ [Hor72; Gaz79]. In a sentence containing a scalar term, such as "I drank some of the beer in the fridge", replacing the term *some* with one higher on the same scale produces an informationally stronger statement: "I drank all of the beer in the fridge", for instance. If informativity corresponds to entailment in the obvious way, then

using a term low on a scale implicates that the terms higher on the scale do not apply, since otherwise some alternative utterance would be more informative and the speaker should have used it.

[Roo03a] points out (pg. 1176) that this formulation, amenable though it is to formal treatment, ignores Grice's original phrase "for the current purposes of the exchange". The result is that some quantity implicatures cannot be treated in this way. Van Rooij cites the following example, due to [Hir85]:

(4) The setting is a job interview.
   a. Interviewer: Do you speak Portuguese?
   b. Applicant: My husband does.

The applicant's answer clearly implicates that she does not speak Portuguese, by the quantity maxim in its original form. But "The applicant speaks Portuguese" and "The applicant's husband speaks Portuguese" do not stand in any kind of entailment relation, so by strict standards of *semantic* informativity (underlying the systems of Horn and Gazdar) this conclusion cannot be drawn. VSI provides a generalisation of semantic informativity: it treats the scalar cases in the same way, but it also puts an ordering on utterances like these, that do not stand in any entailment relation with each other. The applicant's utterance is *less relevant*, by VSI, than the alternative "I speak Portuguese", and so by parallel reasoning to the scalar case it implicates that the alternative does not hold.

4.2.2 · *Mention-some questions [Roo03b]*

What counts as 'fully answering a question' depends not just on the semantics of the question and the answer, but also on the purposes with which it was asked. A classic distinction is between the questions "Who was at the party?" and "Where can I buy an Italian newspaper?" Naming a single newspaper stand is sufficient answer to the second question, while naming a single person who attended the party usually will not satisfy the person asking the first. It is not enough to draw a *semantic* distinction between partial and complete answers, because the same questions can be asked with the acceptability pattern of partial answers reversed:

(5) a. I'm making a survey of newspaper stand quality around the city. First question: where can I buy an Italian newspaper?
   b. I need to hear what happened from somebody who was there. Who was at the party?

What these additional sentences do is make explicit a particular decision problem the questioner is trying to solve. For some problems the VSI of a partial answer is just as high as that of a complete answer (if you tell me I can get an Italian newspaper at the station, I don't care whether I can also get one

at the town hall), while for others the value rises with the specificity of the answer ("All right, John and Pete were at the party, but what about Joan and Petra, you haven't told me about them yet.").

The pragmatic relevance of these facts is the well-known phenomenon of EXHAUSTIFICATION of answers. If I ask "Who (from our circle) was at the party?" and get the answer "John and Pete", I will typically conclude from that "... and not Joan or Petra", by the same kind of quantity-based reasoning mentioned above. In contrast, "Where can I buy an Italian newspaper" is a MENTION-SOME question which does not trigger such exhaustification: there is no temptation to augment the answer "At the station" with the quantity implicature "... and nowhere else". If the inferences are the result of decision-theoretic relevance reasoning, then this phenomenon is entirely to be expected (as is the reversal in (5) above). What matters is not the semantic meaning *simpliciter*, but the influence of that meaning on the decision problem being solved; if more specific answers would be relevant but are not given, the speaker must not be in a position to give them (quantity implicature); alternative answers which would be *no more relevant*, however, as in the case of mention-some questions, have no influence and induce no implicatures.

4.3 · *Calculating VSI with unawareness*

The key to these analyses is to calculate the value of the utterance against the decision problem at hand, and then compare it to the value of putative alternatives. In just the same way, we can calculate the value of an utterance that leads not just to a change in the agent's information but to a change in her awareness. The definition for the more general form is almost exactly the same: we simply remove the subscript for belief update ($[\varphi]$ replaces $[\varphi]_{\mathrm{b}}$), meaning that general updates with an awareness component are allowed. This small change has rather significant conceptual ramifications, though: we are no longer dealing with *information* (propositions whose truth in the actual world is assumed) but with *epistemic change* (which might not be appropriately evaluated in terms of truth and falsity). For this reason (which I will expand on in Section 6.1) [FJ08] called this measure the VALUE OF EPISTEMIC CHANGE (VEC): to emphasise the conceptual gulf between the two (formally almost identical) formulations.

DEFINITION 5.12: Value of epistemic change (VEC).

$$\mathrm{VEC}_D(\varphi) =_d \mathrm{EU}_{D[\varphi]}(\mathrm{BA}(D)) - \mathrm{EU}_{D[\varphi]}(\mathrm{BA}(D[\varphi])).$$

The remarks about properties of VSI of course apply equally to VEC: an utterance is relevant (has non-zero VEC) just if it leads to a change in the set of best actions. For information this is nothing more than the standard account, but for pure awareness updates (triggered by *might*, questions, and so on) we

can note some interesting generalisations about the three kinds of updates given in (3).

One point to note in considering these examples is the joint epistemic status of the decision problem. The simplest possibility is that it is common knowledge between the speaker (who proposes the update) and the hearer (whose situation the decision problem directly represents). In that case the value of an utterance can be simple read off from the decision problem; however most likely the more common situation is when the speaker has imperfect knowledge of exactly which decision problem the hearer is trying to solve. In particular, speaker and hearer might have common knowledge of the probability distribution over states[14] but the utilities of the hearer might be unknown to the speaker, at least in fine detail. This is particularly important because of the 'hard-edged' nature of the VEC calculation. As we will see, "You could take the bicycle" is only VEC-relevant if taking the bicycle is a best action; but then one might ask, why does the speaker not say "You *should* take the bicycle"? If she suffers some uncertainty about precisely which decision problem the hearer is trying to solve, however, the mild hedging becomes much more understandable. With this in mind, let us proceed to the examples.

4.3.1 · *Did you think of taking the bicycle?*

The most typical effect of this update on the decision problem will be nothing more than to broaden the range of actions available at each (proper) state, while the states themselves stay the same. It is easy to see that the only effect this could have on the set of best actions (if indeed the states are totally unchanged) is to add the action b (take the bicycle) to that set (possibly even replacing the current best actions entirely). That is, as we started the chapter by noting, "Did you think of taking the bicycle?" is relevant according to this measure just in case taking the bicycle is a best action: the question can be interpreted as *advice*.

In line with this prediction, mentioning an action which obviously will not be taken is intuitively uncooperative behaviour: "You could rent a limousine," or "Did you consider building an airship?" are not helpful contributions to a conversation about getting to work on time in the morning.

4.3.2 · *There might be a tram strike*

This update will either produce new states or split existing ones (overturning an assumption or increasing the finegrainedness of the agent's beliefs). Overturning an assumption will only produce a change in the best action set if there is sufficient probability mass in the new state to exert an influence; "There might be a tram strike" is as irrelevant as "There might be a military coup

[14] In itself a rather extreme abstraction, but the assumption of a common prior is widespread in the literature on multi-agent systems with probabilistic belief, since results are much harder to find without such constraining structure.

seizing control of the bus service" unless the prior probability of such a strike is relatively high. 'Out of the blue' the statement indeed seems irrelevant, but it would be quite reasonable in a climate of ongoing union disputes, or if the speaker knows that there will be a strike sometime this month but has forgotten exactly when. (Pragmatic *reasoning* based on a presumption of relevance leads to the implicature that a tram strike must be relatively likely, according to the speaker. We will come to this kind of reasoning in Section 5.)

According to the VEC measure, just splitting states can never be relevant. This is because the expected utility of an action in the state is nothing but a weighted average across the worlds in that state, which are the same worlds that would appear in the more finegrained states arrived at by splitting. As I mentioned above, though, a pure finegrainedness update only seems reasonable in a context where the distinction being drawn is indeed 'small' (in terms of outcomes and utilities); in this case a splitting update could be called 'splitting hairs' and seems intuitively irrelevant.[15]

4.3.3 · *If you take the bicycle you might get a flat tyre*

Assuming the consequent of the 'outcome conditional' $a \boxdot\!\!\rightarrow \varphi$ is only possible if the action a is taken (as in this example), only the utility of a (among all the actions) can be affected by the update. That is to say, and remembering that non-zero VEC requires a change in the set of best actions, such an utterance must be an argument *for* or *against* the action being mentioned ('for' if it is not currently a best action, 'against' if it is). This certainly accords with intuition.

More problematic is the possibility that a relevant consequent is introduced by an irrelevant antecedent, as in "If you take the bicycle there might be a tram strike." So long as the tram strike possibility is relevant (see above), this infelicitous conditional will get positive VEC. However I don't think this is a problem for the account, per se: we just need more than a measure of relevance (even one that incorporates the quantity maxim) to correctly predict pragmatic felicity. In particular, the Gricean maxim of manner seems applicable here (since the relevance judgement rests entirely on the consequent, why is the utterance given in such an unnecessarily complicated manner?). It might also be possible to adapt standard Gricean accounts constraining the use of conditionals with unrelated antecedent and consequent to the awareness context.[16]

[15] In a multi-agent setting with several conversational participants, a 'hair-splitting' update might be an invitation to another agent to share information: Ann distinguishes between p and not-p in the hopes that Bob will tell her which is the case. Such examples are only peripherally awareness-related, however, since questions exist for precisely this purpose. Awareness enters the picture only to establish the necessary condition for the question being asked, namely that the person asking it has considered the possibilities involved.

[16] See for example [Fra09, Chapter 5]. The potential difficulty is that such accounts tend to rely on (potential) *informativity*, which is no longer the only measure of relevance in the context of possible awareness updates.

4.4 · *Unawareness and probabilities*

One generalisation applies across all these updates, linking the probabilistic representation with the notion of awareness. I have argued for a notion of assumptions related to expectations of normality; this notion carries over even more strongly into the probabilistic setting. Our assumption-formation faculty (formally represented by the ordering generating the set W_σ) and whatever faculty it is that draws our attention spontaneously to particular atomic formulae (not represented in the system, but implicit in every example in which the agent is already paying attention to certain concepts and not to others) works reasonably well most of the time. It must do, or we would not be able to form stable and reliable beliefs, and go about our daily business more or less successfully.

But if this is the case, then we would expect (as a broad generalisation, not a definite prediction for individual cases) the probability mass hidden by assumptions to be relatively small compared to that 'visible' within the window of awareness. If the actual world is quite likely to be one the agent does not entertain, then her assumption-formation faculty is not doing its job properly. In the Netherlands it is perfectly reasonable to assume that there is no tram strike or military coup in progress; it is evidence of a faulty assumption-formation faculty, on the other hand, to assume that it will not rain without checking a weather forecast.

This carries over to a generalisation about the relevance of awareness updates: when they apply to states of the world (rather than possible actions), the states they draw attention to should generally be associated with extreme utility values. This must be so if the probability mass they reveal is relatively small, since otherwise they stand little chance of changing the set of best actions. And indeed, this prediction too seems to be borne out. "If you cycle you might get a flat tyre" would typically be irrelevant (or, at most, an argument for taking a puncture repair kit), since the probability of a puncture on any particular trip is fairly low. If the agent is in a terrible hurry, though, and the time gained by leaving immediately on the bicycle rather than waiting five minutes for the next tram is significant, the extreme disutility of the puncture (and subsequent lost time) might overturn such a judgement.

5 · *Relevance reasoning*

So far we have only considered evaluating the relevance of utterances 'from the outside', as observers. Our agents too, though, can perform such calculations, and these can form the basis for nontrivial pragmatic reasoning. The general schema is that the hearer assumes the speaker to produce a relevant utterance (by the standard of VEC). If the utterance would only be relevant on the assumption that the speaker possesses some special knowledge, then the hearer concludes that she does indeed possess that knowledge (and typically that she

intended to communicate it). Here is an example.

Example 5.13: Bob the Baker. *Bob (who is an expert baker) is visiting his friend Farmer Pickles (who isn't).*

> PICKLES: *I was going to bake a cake but I haven't got any eggs!*
> BOB: *Did you think of making shortbread instead?*
> PICKLES: *I didn't, in fact I didn't even know that you don't need eggs to make shortbread! Thanks, Bob!*

Remember that raising the possibility of an action can only be an argument *for* that action, under the relevance criterion of VEC. If Bob is advising Pickles to make shortbread, this must be because Bob believes that the recipe for shortbread does not require eggs (since otherwise trying to bake shortbread would be no better than trying to bake a cake, and certainly worse than not baking anything at all). But since Bob is an expert baker (a competence assumption, in the terminology of [RS04]), his belief about shortbread can be taken to be factual.

We have come a long way from the relevance measure of VSI. Bob gives no overt information at all: his utterance is a question, which might indeed alter the common ground (by 'raising issues', or delineating his interest in different potential answers) but certainly does not do so by eliminating worlds. The immediate effect of the question is rather to *add* possibilities, at least in the sense of possible actions — a *reduction* in information as standardly conceived. And yet, by pragmatic reasoning Pickles indeed gains just the information he needs.

Indeed, I would argue that Grice's famous petrol example belongs more in this context than as an example of reasoning from a purely informative utterance. "There is a garage around the corner" is very likely a proposition the car owner was not attending to (otherwise, parallel to the situation with Walt and the car keys, she should be looking around the corner to check). The action "Go around the corner and get petrol from the garage" is equally unlikely to be a part of the original decision problem, but once it has been incorporated the reasoning is entirely parallel.

The example highlights one deficiency of the propositional approach: awareness of the action "Go around the corner and get petrol from the garage" is clearly closely related to awareness of the garage being around the corner, a relation which is entirely obscured by representing them as distinct proposition letters. A first-order model in which the agent may be unaware of actions, properties, and objects and in which these combine in the natural ways would be much more appropriate for such examples; I will consider the beginnings of such a theory in the next chapter. Here is a third example on the same structural lines, due to Anton Benz, which requires a similar approach.

Example 5.14: Travel expenses. *I am to submit a form requesting repayment of travel expenses to the administrators of the* ZAS*. I ask at the front desk where I should go, and am told, "Mrs Schmidt is in room 2.15."*

The reasoning that leads me to deliver my expense form to Mrs Schmidt is exactly parallel to that leading Grice's stranded driver to the garage for petrol. However Benz's example has a further symmetry property which causes difficulties for standard probabilistic accounts, but which can be solved using unawareness.

5.1 · *A digression on symmetry*

An awareness-based account does better than a purely propositional account for this example because we can have actions 'offstage' for the agent, so that the information that Mrs Schmidt is in a particular room brings with it the implicature that a particular action is optimal. However Benz was concerned with a refinement of the example, which raises even tougher questions about the use of possible-world semantics for such problems:

Example 5.15: Symmetrical travel expenses (Benz). *As before, I am to submit a travel expense form at the* ZAS*. This time I know that it must go to either Mrs Schmidt or Mr Müller, and that their rooms are 2.15 and 2.16; I don't know, though, who has which room, nor do I know who should get the form. As before, the front desk assistant tells me, "Mrs Schmidt is in room 2.15."*

The problem for standard theories is the extreme symmetry of the example. Given my background knowledge, the proposition expressed by "Mrs Schmidt is in room 2.15" is identical to the proposition expressed by "Mr Müller is in room 2.16". Under such (admittedly artificial) conditions we have no formal representation of the intuitive 'aboutness' of the two utterances (that the first is about Mrs Schmidt and room 2.15, while the second is about Mr Müller and room 2.16), since both pick out exactly the same subset of my belief set. But of course, this is precisely what a theory of awareness gives us: an utterance is not reduced to its propositional information, but is (in effect) a pair consisting of its propositional meaning and the things it 'talks about' (the atomic formula as a syntactic object, or in our informal extension for this example, the people and places it mentions).

That is not enough to magically solve all our difficulties, however. It is hard to imagine how I could know that Mrs Schimdt and Mr Müller can be found in rooms 2.15 and 2.16 (although in which order I am uncertain) without already being aware of the two people and places. The awareness update produced by "Mrs Schmidt is in room 2.15", then, is vacuous, and it seems we are left in the same quandary as the theories without awareness.

There is, however, a loophole: pragmatic reasoning. The attention update

has no effect, but it may have been *intended* to have an effect. Specifically, it may have been intended to make me aware of Mrs Schmidt just as in the first example, to send me to room 2.15 to hand in my travel expense form. This would be the case if the speaker thought I was unaware of Schmidt and Müller, which seems reasonable, but in fact it also holds for all deeper nestings of uncertainty. In all epistemic situations except common knowledge that I am aware of Mrs Schmidt (that is, any finite-depth mutual certainty followed by uncertainty), a pragmatic case can be made for using "Mrs Schmidt is in room 2.15" to indicate that that is where I should go.[17] In contrast, *no* pragmatic case can be made for using the same statement to direct attention to Mr Müller in *any* configuration of mutual belief or uncertainty. In any case but common knowledge, there exists the possibility that the assistant uses the utterance to send me to Mrs Schmidt; since this is the only possibility that makes the utterance relevant (it can never draw attention to Mr Müller), the mere existence of such a possibility makes it the preferred interpretation.

I don't mean to claim that we go through such complex reasoning in considering such simple examples. Rather, I suggest that this phenomenon lies at the functional root of a far simpler notion: the influence of aboutness on Grice's maxim of manner. The possibility of awareness confusions, even if only in rather extreme cases, justifies a manner 'submaxim': in order to convey information about x, use (preferentially) an expression about x. This is of course nothing but common sense, and it takes an artificial example like the one we are considering here to force us to acknowledge that our formal systems, elegant though they may be, do not represent this level of common sense behaviour. To this extent the awareness model is an improvement: it tells us *why* our common sense behaves the way it does, if not necessarily (in this case) *how*.

5.1.1 · *Horn's division*

Unawareness can help solve such symmetry problems because it is inherently asymmetric. An agent aware of p can understand the viewpoint of *un*awareness of p, but the converse is impossible. So in the previous example, despite the semantic equivalence (given the background information) of the information about Schmidt and Müller, whichever one is not mentioned is 'invisible' to the agent.

Another application of this idea is to Horn's division of pragmatic

[17] The individual cases become more abstruse as the nesting depth increases, but the pattern is very regular. If at level n I am unaware of Schmidt, then at the same level the assistant deliberately draws attention to her. At level $n-1$ I believe that the assistant has deliberately drawn attention to Schmidt, conveying the required message despite my actually already being aware of Schmidt (at $n-1$). Thus the assistant at $n-1$ can successfully apply this tactic: I get the right message, despite knowing that it is for the wrong reasons; at level $n-2$ I hold it possible that this is his intent, despite our mutual knowledge that I am aware of Schmidt, and so on back to level 0.

LABOUR: the rule that when marked and unmarked expressions have the same semantic meaning, the marked expression is used for non-stereotypical cases of the meaning while the unmarked expression goes with stereotypical cases (the notion, although not the term, comes from [Hor84]). This is another case where game theory has produced effective but perhaps over-complicated explanations; [Ros04] shows that Horn's division emerges from certain kinds of evolutionary models (although an equally important part of the paper is given over to game-theoretic models that *do not* produce Horn's division, in some cases quite counter to first expectations).

Van Rooij also discusses static (ahistorical) game-theoretic models, but concludes that they do not suffice to explain the Horn pattern, for reasons that have a lot to do (again) with symmetry. Without getting into the details, the competing possible strategies produced by systematically permuting utterances and meanings have many structural properties in common with the Horn strategy (even though the Horn strategy is in the end the most efficient at the global level), so that standard game-theoretic techniques cannot 'see the differences' between them. Certain kinds of evolutionary model, on the other hand, allow efficiency to influence the long-term behaviour of the model; since the historical dimension is a crucial part of this process, Van Rooij concludes that Horn's division is (in its general form) a *convention* of the language, stabilised over the long term by its relative efficiency.

The same result can, however, be derived using unawareness under simpler (and arguably more natural) assumptions. Gricean pragmatic reasoning proceeds by comparisons between the utterance *actually* made and the possible alternatives the speaker *could* have said. If the awareness relations between these are systematically asymmetric, no appeal to conventions or long-term linguistic evolution need be made.

Suppose that the marked form is one the agent is typically unaware of, and that the marked form triggers awareness of the unmarked form but not vice versa. Van Rooij begins his paper with a standard example: "Miss X produced a series of sounds that corresponded closely with the score of 'Home Sweet Home'." After puzzling over this for a moment, you probably realised that she sang the song (this insight sets in train the pragmatic reasoning that starts with "Why didn't he just *say so* then?"). If he had said instead that Miss X sang 'Home Sweet Home', nothing would have drawn your attention to the possible alternative utterance "Miss X produced a series of sounds…".

The intuitive appeal of this account is its asymmetry. The usual case (unmarked form, stereotypical meaning) involves unawareness of both the marked utterance in its syntactic form and the very dimension of variation on which the stereotypical and unexpected events differ (that is, whether the singing was good or bad). From the perspective of awareness of the marked

form, we can still 'see' this alternative possibility; this corresponds to the intuitive question "If she just sang (in an ordinary way), why didn't he just say so?" Searching for a motivation, we must spontaneously come to attend to the quality of the singing as a possible additional variable.[18] Only after attending to the possibility of bad singing do we realise *explicitly* that our stereotypical assumption about singing is that it is pleasant.

This means that the account differs substantially from Van Rooij's treatment of 'stereotypicality'. For him the stereotypical event is the one with highest probability (this is mathematically required for Horn's division to be the most efficient long-term strategy). While generally appealing, this idea runs into difficulties with specific cases. What if, for instance, Miss X is *always* a terrible singer? Then we would be forced (under the explicit terms of Van Rooij's account) to conclude from the marked expression that she sang *well*, unless we fall back on an ad hoc procedure for generalising probabilities that relates the stereotypical interpretation of the (specific) utterance about Miss X to knowledge about singers in general. That stereotypical singing is pleasant singing is, I would suggest, a fact about *awareness* rather than a fact about probabilities. (The lowest singing-worlds in our $\preceq$ orderings tend to be pleasant-singing ones.) This fact is *connected* to frequencies of events, in that our awareness orderings must behave roughly consistently with probabilistic expectations, but it does not reduce to them. And, in particular, if we are not attending to the question of quality, we may very well hold a general assumption (that the singing was pleasant) which we would repudiate when the question is explicitly raised (bearing in mind what we also know about Miss X).

6 · *Summary*

The treatment of Horn's division of pragmatic labour shows one way in which unawareness can aid formal pragmatics: in simplifying the representations of core phenomena, reducing complex game-theoretic explanations to more tractable decision theory. More generally, awareness can reduce the unwanted effects of symmetry, whether probabilistic and numerical (as in the case of Benz's expense form) or structural (in the game-theoretic representation of Horn's division).

A second achievement is the extension of standard techniques (such as relevance reasoning from VSI) beyond the standard cases (the shift from informative to allusive dialogue). Formal pragmatics has had its greatest successes in reasoning based on simple informative language use, since semantic meaning gives an immediate formal handle with which to manipulate utterances in rea-

[18] My account is silent on how this particular variable is arrived at; but so indeed are all others that I am aware of. Game-theoretic approaches assume that the possible meaning space is already given, so that the only possibilities are "sang badly" and "sang well" (why not also "whistled", for instance?). Informal accounts of course fare no better, on what is essentially an awareness question.

soning. Unawareness provides one way in which the 'tyranny of informativity' can be broken; certainly this is only a small part of non-informative language use, but any movement in this direction can only be to the benefit of the field. In particular, unawareness models explain how questions, statements with *might*, and similar paradigms of uninformativity can be pragmatically interpreted as advice, against all semantic expectation.

In the other direction, decision-theoretic models have told us something interesting about the pragmatic conditions under which drawing attention to new possibilities counts as cooperative language use. In particular, the requirement that an awareness update make a difference to the actions the agent takes as best leads to a general condition on the relation between assumptions and probabilities: if an assumption is overturned then either it was concealing significant probabilistic weight (the 'head-slap moment' that greets Pat's helpful suggestion to Walt) or, in the more usual case where the concealed probability mass is relatively small, the utilities of outcomes involved must be extreme (whether extremely large or extremely small).

Let me finish by mentioning a persistent loose end, which raises questions about the interpretation of the notion of 'information' in models with awareness.

6.1 · *Some speculation: hurting attention?*

The problem is that not all of the arguments for VSI transfer immediately to the unawareness context. In particular, while I am convinced that information must have a non-negative value (and that putative counterexamples rest on an insufficiently serious interpretation of probabilistic expectations), the parallel argument for attention seems to me rather weaker.

The main difference is that information converges in the limit, while awareness updates can broaden the agent's horizons apparently without restraint. Information has non-negative value because each informative update brings the agent closer to the fixpoint of no uncertainty (a belief state containing only the actual world); but each awareness update takes her *further away* from such a fixpoint, by introducing new possibilities.

Consider again the unfortunate case of Jim and Joan and Jane. A malevolent speaker who saw Jane and Joan in the café could manoeuvre Jim into making a wrong decision, by saying "Joan is in the café". But there are limits to his potential for mischief: the more information he chooses to give, the more appropriate Jim's reaction will be, so that in the limit Jim knows which the actual world is and takes the action maximising (not expected but actual) utility. A malevolent speaker in the system with awareness, on the other hand, can cause unending misery, because outré possibilities can always be found that are highly unlikely but that carry extreme utilities.

This brings us to considerations of scepticism, which I am not sure form a genuine continuum with mundane awareness data (see the discussion in

Chapter 7 Section 2.1). In any case the pragmatic reasoning we have been concerned with in this chapter makes the Gricean assumption of cooperation, so the problematic consequence of malignancy is irrelevant; I wonder, though, whether similar problems will arise in a setting of only partial coordination (see the treatment of credibility in [Fra09]) and what the consequences for an awareness-relative measure of relevance should be.

Chapter 6

Data semantics with unawareness

> Facts alone are wanted in life. Plant nothing else, and root out everything else. You can only form the minds of reasoning animals upon Facts: nothing else will ever be of service to them.
>
> Gradgrind in Charles Dickens's *Hard Times*

The final model I will present in this dissertation is a curious sort of hybrid. It takes aspects of the object-based awareness semantics of Board and Chung [BC07], a 'gap semantics' for assumptions as implicitly suggested by the subjective state-space approach of Heifetz, Meier, and Schipper [HMS06], and a treatment of information and inference from Frank Veltman's data semantics [Vel81]. Let us focus first on the awareness part of the picture, and the motivation for object-based assumptions; we will come back to data semantics in a moment.

I spoke briefly in the introduction about how the subjective state-space models suggest a 'gap semantics' for object-based unawareness: in a sub-space where the object a is not part of the vocabulary, a simply does not appear in the models forming 'worlds' or points in the space. This seems perfectly natural for unawareness of *objects* (as opposed to propositions), and raises interesting quantificational possibilities; if the agent is unaware of any object with property P, for instance, the models in her assumption set will all validate $\neg\exists x : P(x)$.

The immediate motivation for such an idea comes from scenarios like Walt and his keys, example 1.2 from Chapter 1. Here is a modified version of the example:

(1) Walt is looking for his car-keys. He has in mind two places they might be: on the kitchen table, or in a basket in the bookcase. He checks the kitchen table and finds no keys.
   a. Walt: They must be in the basket.
   b. Pat: Did you check your pockets?
   c. Walt: Good point, they may not be in the basket after all.

Walt's conclusion in (1-a) is only defensible if he assumes that the keys are not

in any of the multitude of places that he does not have in mind. A natural way to represent this is to take models in which the worlds have variable domains: some worlds include all the objects but some are SMALL WORLDS whose domains contain a proper subset of the objects that 'really' exist. Walt's assumption of absence comes down to him filling his belief set with small worlds: those worlds at which only the objects he is aware of exist (and all natural laws are still followed). If only two places exist where the keys might be (in the small worlds of Walt's belief set), when one is eliminated the other must be the correct location. This also gives the right results when Pat makes him aware of further alternatives: he gives up his conclusion, since his belief set must shift to worlds with three-element domains where it no longer holds.

However it is not only becoming aware of explicit alternatives that can cause Walt to give up his conclusion. Being confronted with contradictory evidence will do so as well.

(2) As before, Walt entertains two possibilities for the keys; he checks the kitchen table and does not find them.
   a. Walt: They must be in the basket on the bookcase. [*He checks.*]
   b. Walt: They're not! But they must be *somewhere* . . . Where could they be?
   c. Pat: Did you check your pockets?
   d. Walt: Ah, what an idiot I am, that must be where they are.

In the models of the previous chapters, after checking beside the basket Walt would arrive at the absurd state: his belief set would contain no worlds at all. I have suggested that arriving at the absurd state is less problematic in models with unawareness than those without, since the agent can always imagine that there must be some factor that he is unaware of, but until now this explanation has remained entirely informal. While the agent must give up *some* assumption to escape his predicament, so long as assumptions are propositional (rather than object-based) it is formally unclear how to select the correct assumption to overturn, and intuitively unclear why the agent should become aware of just the right possibility to solve his difficulties. Matters are much simpler in the case of object-based unawareness: clearly all that is needed is *more objects*, and Walt need not even become aware of them, so long as he is able to believe that there exists something that he is unaware of.

So much for the unawareness models. What is the connection with data semantics? The answer is that these examples turn on the distinction between (immediate) *data* and (indirect) *conclusions*. Recall that in the models of previous chapters, becoming aware of new possibilities can overturn previous beliefs: after learning that p and then becoming aware of q, the agent need not still believe that p. I have argued that this is not entirely unreasonable when the

updates represent incoming *linguistic* information: the speaker, too, might have been unaware of q when she announced p, and might repudiate her earlier certainty once she has become aware of the new possibilities. The case is rather different when agents go out and discover information for themselves, however, as when Walt checks in the basket and sees that the keys are not there. We most definitely do not want him to come to doubt this information when he becomes aware of the possibility that the keys might be in his pocket! It seems that we must distinguish between Walt's immediate data (that the keys are not on the kitchen table) and the conclusions he draws from that data (that they must be in the basket).

Data semantics was designed exactly to make this distinction, although of course without unawareness in mind. The idea is, very roughly, that an agent's data defines a partial model, while her conclusions are whatever holds in all possible completions of this model. Data, that is, is 'definite', while conclusions are the result of quantification over a set of possible extensions of the data. In the original theory these possible extensions are fixed within a given model, but the natural way to incorporate unawareness seems to be to allow the set of possible extensions to expand as the agent becomes aware of new possibilities. This is indeed the tack we will take.

The rest of the chapter is arranged as follows. In the next section I introduce the theory of data semantics, and give a naïve attempt at adapting this theory for awareness using the gap semantics. The attempt fails, of course (otherwise I would hardly refer to it as "naïve"), but its failure makes clearer what we need from the system. Following this I construct a model that combines data semantics with a 'small world' representation of assumptions of absence.

1 · *Data semantics*

Data semantics is a semantic theory first published by Frank Veltman in the year I was born, and elaborated by him and Fred Landman in the decade or so following [Vel81; Lan84; Vel85; Vel86; Lan86]. The intent of the theory is to explain what it should mean for a statement to be "true on the basis of the available evidence". This motivation already suggests a fruitful combination with unawareness: our beliefs about what is true on the basis of some piece of evidence might shift quite radically without the evidence itself changing, if we take new possibilities into account in the way an awareness model suggests.

The formal apparatus of data semantics has appeared in several guises. Perhaps the most familiar-looking is that given in [Vel86], in which "the available evidence" is given by a partial propositional valuation. [Vel81] instead sees the available evidence as a set of (possible) FACTS.[1] I will give the presentation

[1] I will use the term rather loosely, since I am not concerned with the extensive philosophical debate on precisely what facts are (and what facts there are). In all fact-based data semantics models "fact" means *possible* fact (which may not actually obtain). For Veltman in 1981 the conjunction of a set of

in terms of data sets, since we will want first-order rather than propositional valuations in the end.

Definition 6.1: Structures for data semantics. *A* data lattice *is a meet-semilattice with a least element:* $\mathfrak{L} = \langle \mathcal{F}, \circ, \bot \rangle$*, where*

- $\mathcal{F}$ *is the set of possible facts;*
- $\bot \in \mathcal{F}$ *is the absurd, contradictory, impossible fact;*
- $\circ$ *is a meet operation on* $\mathcal{F}$ *(commutative, associative, and idempotent);*
- $\bot$ *is the least element: for all* $f \in \mathcal{F}$*,* $f \circ \bot = \bot$*.*

Given such a lattice $\mathfrak{L}$ *we can define an information ordering* $\leq_{\mathfrak{L}}$ *on* $\mathcal{F}$*:*

$$f \leq_{\mathfrak{L}} g \text{ iff } (f \circ g) = f,$$

and $f \leq_{\mathfrak{L}} g$ *can be read as "*f *includes all the information* g *includes (and possibly more)" (*f includes g*);* $\bot$ *is the least element in this order (for all* $f \in \mathcal{F}$*,* $\bot \leq_{\mathfrak{L}} f$*). We can also define a* partial *meet operation* $\&_{\mathfrak{L}}$ *on* $\mathcal{F} \setminus \bot$*:*

$$f \,\&_{\mathfrak{L}}\, g = \begin{cases} f \circ g & \text{if } f \circ g \neq \bot, \\ \textit{undefined} & \textit{otherwise.} \end{cases}$$

Two facts f *and* g *are* incompatible *in* $\mathfrak{L}$ *if* $f \circ g = \bot$*; I also write this* $f \perp g$*.*

Intuitively $\circ$ represents conjunction; the representation with & is perhaps more philosophically respectable for anyone who objects to the notion of an 'impossible fact' but is somewhat less convenient for defining data. The lattice as a whole represents what *could* be the case; an agent's data represents those facts that he is acquainted with, in other words those that he has personally and experientially verified.

Definition 6.2: Data. *Let* $\mathfrak{L} = \langle \mathcal{F}, \circ, \bot \rangle$ *be a data lattice. A* filter *on* $\mathfrak{L}$ *is a set* $\mathfrak{d} \subseteq \mathcal{F}$ *such that* $f \circ g \in \mathfrak{d}$ *iff* $f \in \mathfrak{d}$ *and* $g \in \mathfrak{d}$*. A filter* $\mathfrak{d}$ *is* proper *if* $\bot \notin \mathfrak{d}$*. A* data set *for* $\mathfrak{L}$ *is a proper filter on* $\mathfrak{L}$*.*

The data lattices I am concerned with in this chapter are finite. This has two consequences of importance. Firstly, that any data set $\mathfrak{d}$ (indeed any filter,

possible facts counts as another possible fact; Landman in 1982 (when [Lan84] was written) uses the term only for what we might call 'atomic' facts ("[T]here is no fact which expresses exactly the information that contains the information expressed in *it rains* and *it snows*." pg. 169); and Veltman by 1986 has given up on facts altogether. For all versions of the theory it matters, however, that there is a distinction between facts (or possible data) and propositions: not all propositions are possible facts (not all propositions can be data).

proper or not) is generated by a single fact f: $g \in \mathfrak{d}$ iff $f \leq g$. Secondly, that the lattices contain ATOMIC FACTS; a fact is atomic if it is maximal in the $\leq$ ordering. The atomic facts will correspond to atomic sentences in our logical language. If we diagram a lattice with $\leq$ increasing up the page, then a data set corresponds to an 'upwards cone' (although it must also be closed under $\circ$, see Figure 6.1).

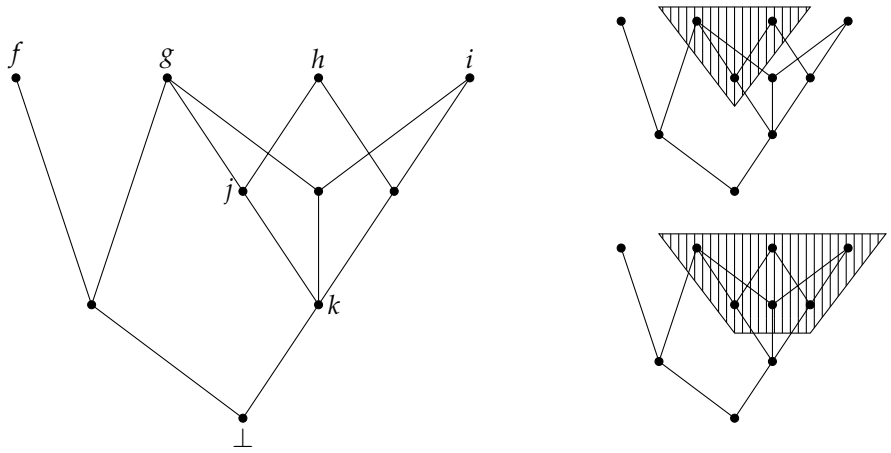


Figure 6.1: A data lattice, a data set, and a set of facts that is not a data set. Lines indicate the relation $\leq$, increasing upwards. The meet of two facts is its greatest lower bound; for example $j = g \circ h$, and $\bot = f \circ j$. (This means that f and j are incompatible.) The facts f, g, h, i are atomic. In the upper right figure the shaded area is a data set; the lower right figure is *not* a data set: it contains (for example) j and i but not $k = i \circ j$.

Truth "on the basis of evidence" gives rise to a three-valued logic: the evidence may support φ, contradict φ, or not decide the issue. I cannot simply give the definitions that Veltman and Landman use, however, since they do not consider first-order languages: their models do not support quantification. To interpret the universal quantifier I need to equip my data lattices with a domain of objects. I will also make two simplifying assumptions that are philosophically completely unjustified but that make the models much more convenient to work with.

Firstly, I am going to assume that names of objects are rigid, and that each object has exactly one name. The domain of objects itself 'stands in' for the set of object constant symbols (names of objects) in the logical language: Pa says that a has the property P, not that the-interpretation-of-a has the property. This is purely for expediency: it lets me talk about the agents being aware of particular objects rather than the names of those objects. Secondly, I assume

that the atomic facts have a particular structure: that negative atomic facts exist. For each atomic fact f there exists a particular atomic fact g such that $f \perp g$; if f stands for the object a having the property P then g stands for *a not* having the property (and vice versa). I will make a third simplification also: I give definitions only for predicates, rather than including relations of higher arity. (Unlike the first two restrictions, this is simply to keep the definitions concise; the extension to the relational case presents no conceptual problems whatsoever.) With those provisos out of the way, here are the definitions.

Definition 6.3: Syntax. *Let Ω be a vocabulary of predicate symbols, and D a set of objects. The language $\mathscr{L}^{\Omega,D,\forall,\Diamond}$ is a first-order language with quantifiers $\forall, \exists$ (and a countable set of variables for them to bind), non-logical vocabulary Ω, object constants from D, and modal operators must and may .*

I follow [Vel81] in using *may* rather than *might*; it is nonetheless meant (both here and in Veltman's work) epistemically rather than deontically.

Definition 6.4: First-order data model. *Fix Ω and D as above. A* first-order data model *for $\mathscr{L}^{\Omega,D,\forall,\Diamond}$ is a triple $\mathfrak{M} = \langle \mathfrak{L}, D, \mathcal{I} \rangle$ where*

- *$\mathfrak{L} = \langle \mathcal{F}, \circ, \perp \rangle$ is a data lattice;*
- *D is the domain of objects that exist;*
- *$\mathcal{I} : \Omega \times \{+, 1\} \times D \to \mathcal{F}$ is an interpretation function giving positive and negative atomic facts: $\mathcal{I}(P, +, a)$ gives the atomic fact "that a has property P" while $\mathcal{I}(P, -, a)$ gives the atomic fact "that a does not have property P"; and*
- *for every predicate $P \in \Omega$ and object $a \in D$, $\mathcal{I}(P, +, a) \circ \mathcal{I}(P, -, a) = \perp$.*

Now we can give the truth definition. A formula φ (which may contain free variables) is evaluated against a model $\mathfrak{M}$, a data set $\mathfrak{d}$ from that model, and an assignment s of variables to objects in the domain. The formula may be verified by the data (written $\mathfrak{M}, \mathfrak{d}, s \models \varphi$), it may be falsified by the data (written $\mathfrak{M}, \mathfrak{d}, s \dashv \varphi$), or the matter may not be decided (on the basis of that data).

Definition 6.5: Semantics. *Let $\mathfrak{M} = \langle \mathfrak{L}, D, \mathcal{I} \rangle$ be a first-order data model for $\mathscr{L}^{\Omega,D,\forall,\Diamond}$. Let $\mathfrak{d}$ be a data set for $\mathfrak{L}$.*

Atomic formulae:

$\mathfrak{M}, \mathfrak{d}, s \models Pa$ *iff* $\mathcal{I}(P, +, a) \in \mathfrak{d}$ *(where $a \in D$ is an object)*
$\mathfrak{M}, \mathfrak{d}, s \models Px$ *iff* $\mathcal{I}(P, +, s(x)) \in \mathfrak{d}$ *(where x is a variable)*
$\mathfrak{M}, \mathfrak{d}, s =\!| Pa$ *iff* $\mathcal{I}(P, -, a) \in \mathfrak{d}$ *(where $a \in D$ is an object)*
$\mathfrak{M}, \mathfrak{d}, s =\!| Px$ *iff* $\mathcal{I}(P, -, s(x)) \in \mathfrak{d}$ *(where x is a variable)*

Conjunction, disjunction, negation:

$\mathfrak{M}, \mathfrak{d}, s \models \varphi \vee \psi$ *iff* $\mathfrak{M}, \mathfrak{d}, s \models \varphi$ *or* $\mathfrak{M}, \mathfrak{d}, s \models \psi$
$\mathfrak{M}, \mathfrak{d}, s =\!| \varphi \vee \psi$ *iff* $\mathfrak{M}, \mathfrak{d}, s =\!| \varphi$ *and* $\mathfrak{M}, \mathfrak{d}, s =\!| \psi$
$\mathfrak{M}, \mathfrak{d}, s \models \varphi \wedge \psi$ *iff* $\mathfrak{M}, \delta, s \models \varphi$ *and* $\mathfrak{M}, \mathfrak{d}, s \models \psi$
$\mathfrak{M}, \mathfrak{d}, s =\!| \varphi \wedge \psi$ *iff* $\mathfrak{M}, \delta, s =\!| \varphi$ *or* $\mathfrak{M}, \mathfrak{d}, s =\!| \psi$
$\mathfrak{M}, \mathfrak{d}, s \models \neg\varphi$ *iff* $\mathfrak{M}, \mathfrak{d}, s =\!| \varphi$
$\mathfrak{M}, \mathfrak{d}, s =\!| \neg\varphi$ *iff* $\mathfrak{M}, \mathfrak{d}, s \models \varphi$

Quantification ($s[x/a]$ is the variable assignment just like s but mapping x to $a \in D$; $\varphi[x/a]$ is the formula just like φ but with all free occurrences of x replaced by a):

$\mathfrak{M}, \mathfrak{d}, s \models \exists x : \varphi$ *iff for some* $a \in D : M, \mathfrak{d}, s[x/a] \models \varphi[x/a]$
$\mathfrak{M}, \mathfrak{d}, s =\!| \exists x\varphi$ *iff for all* $a \in D : \mathfrak{M}, \mathfrak{d}, s[x/a] =\!| \varphi[x/a]$
$\mathfrak{M}, \mathfrak{d}, s \models \forall x\varphi$ *iff for all* $a \in D : \mathfrak{M}, \mathfrak{d}, s[x/a] \models \varphi[x/a]$
$\mathfrak{M}, \mathfrak{d}, s =\!| \forall x\varphi$ *iff for some* $a \in D : \mathfrak{M}, \mathfrak{d}, s[x/a] =\!| \varphi[x/a]$

Modals:

$\mathfrak{M}, \mathfrak{d}, s \models$ *may* φ *iff there is some data set* $\mathfrak{d}'$ *for* $\mathfrak{L}$ *such that* $\mathfrak{d}' \supseteq \mathfrak{d}$ *and* $\mathfrak{M}, \mathfrak{d}', s \models \varphi$
$\mathfrak{M}, \mathfrak{d}, s =\!|$ *may* φ *iff there is no data set* $\mathfrak{d}'$ *for* $\mathfrak{L}$ *such that* $\mathfrak{d}' \supseteq \mathfrak{d}$ *and* $\mathfrak{M}, \mathfrak{d}', s \models \varphi$
$\mathfrak{M}, \mathfrak{d}, s \models$ *must* φ *iff there is no data set* $\mathfrak{d}'$ *for* $\mathfrak{L}$ *such that* $\mathfrak{d}' \supseteq \mathfrak{d}$ *and* $\mathfrak{M}, \mathfrak{d}', s =\!| \varphi$
$\mathfrak{M}, \mathfrak{d}, s =\!|$ *must* φ *iff there is some data set* $\mathfrak{d}'$ *for* $\mathfrak{L}$ *such that* $\mathfrak{d}' \supseteq \mathfrak{d}$ *and* $\mathfrak{M}, \mathfrak{d}', s =\!| \varphi$

As usual we write $\mathfrak{M}, \mathfrak{d} \models \varphi$ *when* $\mathfrak{M}, \mathfrak{d}, s \models \varphi$ *for every assignment s, and likewise for* $=\!|$.

The clause for atomic formulae includes my two simplifying assumptions: objects stand in for their own names, and *Pa* is *falsified* if the negative atomic fact $\mathcal{I}(P, -, a)$ is in the data set. The clauses for conjunction, disjunction, negation, and the modals are standard (they pass through the variable assignment

unchanged); the quantification clauses ask for witnesses from the domain D as you would expect.

[Vel81] defines a notion of stability: a formula is T-stable if 'once true it stays true' (if $\mathfrak{d}$ verifies φ then any $\mathfrak{d}' \supseteq \mathfrak{d}$ also verifies φ) and is F-stable if 'once false it stays false' (if $\mathfrak{d}$ falsifies φ then any $\mathfrak{d}' \supseteq \mathfrak{d}$ also falsifies φ). Formulae that do not contain modals are both T- and F-stable but *may* φ is not T-stable and *must* φ is not F-stable.

A formula *may* φ is true at a data set $\mathfrak{d}$ if $\mathfrak{d}$ does not exclude the possibility that φ. Suppose we are investigating a crime; at a preliminary stage of the investigation we might say "According to the evidence so far, the butler may have done it". It is perfectly reasonable, however, for additional evidence acquired later to rule out this possibility: *may* φ goes from being true (on the basis of the data) to being false (on the basis of *more* data). Formulae with "*must*", on the other hand, remain true under data extension but are not F-stable. While the possibility exists that the butler is innocent, it is *false* (not undefined) that he "must have done it". Growing evidence against him, however, may naturally overturn this judgement: when his fingerprints are discovered all over the butter dish, "oh, then he must have done it" becomes true not false.

Note, however, that this is *not* the kind of non-monotonicity on display in examples (1) and (2). Under these semantics *may* is F-stable (once false it remains false) and *must* is T-stable (once true it remains true). In example (1) "The keys must be behind the door" changes from true to false, and if we think of Walt's *implicit* beliefs (under his assumptions) then "The keys may be in my pocket" changes from false to true.

Data semantics makes *must* φ weaker than φ, in that there are states that support *must* φ but do not (yet) support φ. (This is of course the reverse of the more usual modal pattern, in which a necessity modal is *stronger* than the non-modal equivalent sentence.) The essence of the distinction is the notion of 'immediacy' of consequence. We can claim "φ" if φ follows 'immediately' or 'directly' from our data; if our data makes it certain that φ but only 'indirectly' then we must say "*must* φ". And it is this distinction that is at the root of the non-monotonicity that I want to add: indirect consequences come from *quantification* over a set of possibilities (those that appear in the data lattice), and awareness can naturally change this set. Direct consequences, on the other hand, are 'the data itself' and do not change as the agent's awareness of possibilities expands.

In other words: the agent's assumptions should restrict the data lattice of which her data is a subset. The less possibilities are included in the lattice, the more conclusions he will draw under *must* (conclusions like "The keys must be in the basket (because they are not on the table)"). The trick will be to define the *right* restrictions...

1.1 · *A first attempt*

The most obvious way to include assumptions of absence into this scheme is simply to omit any facts that make reference to the objects the agent is unaware of. Example (2) adds an extra complication, but in fact already (1) is enough to show that this strategy will not work. Seeing why will make much clearer what the successful strategy must be, however.

Here is a schematic picture of how we would go about it. Suppose $\mathfrak{M} = \langle \mathfrak{L}, D, \mathcal{I} \rangle$ is a first-order data model and we want to restrict it to a domain $D' \subseteq D$. We need to remove from $\mathcal{F}$ every fact that is about some element of $D \setminus D'$. What does "aboutness" mean here?

Certainly an atomic fact $\mathcal{I}(P, +, a)$ (or $\mathcal{I}(P, -, a)$) is about a. And the data lattice encodes an idea of information containment: if $f \leq_{\mathfrak{L}} g$ then f 'already contains' the information in g, so if g is about a then so is f.[2] If $f \leq I(P, +, a)$ then f already contains the information in $I(P, +, a)$ — information which is about a. Let us assume that the *only* way for a fact to be about a is via some atomic fact about a. Then our restricted set of facts will be given by

$$\{f \in \mathcal{F} \,;\, \text{there is no } P \in \Omega \text{ and } a \in D \setminus D' \text{ such that } f \leq I(P, +, a) \text{ or } f \leq I(P, -, a)\}.$$

That is rather a mouthful, but a schematic picture is fairly easy to draw. Figure 6.2 overleaf represents the *entire data model* under the assumption of absence that our restriction is intended to represent.[3]

So let us try to apply this picture to Walt in example (1); the model is overleaf in Figure 6.3. For simplicity we'll restrict the actual possibilities to three places: table, in the basket, and his pockets. For even more simplicity, I will pretend that "containing the keys" is a property: keys(b) says the basket has the keys, and so on. This is purely for convenience in drawing the diagram: the only atomic facts I include are those regarding the location of the keys, and I'll give them some intuitive names (we will be working with variations on this scenario for the rest of the chapter). The atomic fact that the keys are in the basket ($\mathcal{I}(\text{keys}, +, b)$) is b^+; that they are not in the basket is b^- (remember that we *do* have negative *atomic* facts), and likewise for on the table (t^+, t^-) and in

[2] Note that our semantics does not include universal facts, which would otherwise upset this argument. When data supports $\forall x \varphi$ it is not because some particular fact supports this but for the usual quantificational reason that every substitution into φ is (individually) supported.

[3] According to the proposed definition of 'aboutness', $\bot$ is about *every* object and would be removed by any restriction. I apply the construction above instead to the 'partial meet' view of the lattice, in which two facts are incompatible if they do not have a meet in $\mathcal{F}$. [Lan84] uses De Morgan lattices to represent the propositional structure induced by a set of facts; in a De Morgan lattice there can be several different contradictions (and several different tautologies), each 'about' a different set of atomic facts. This is yet another representation of partiality; I have not explored its appropriateness for representing unawareness.

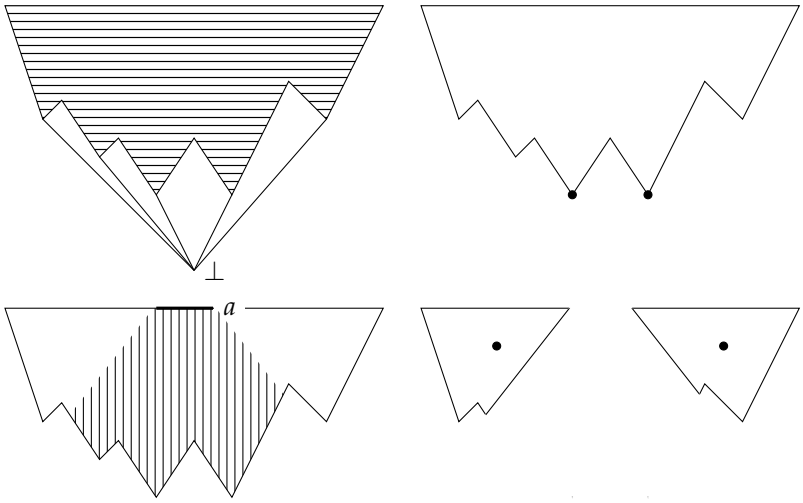


Figure 6.2: Schematic view of a data lattice restricted by assumptions of absence. The top-left figure shows the entire lattice; the shaded portion contains the 'genuinely possible' facts that are not $\perp$ (with the structural details omitted). The next figure shows the 'partial meet' view of such a lattice: without $\perp$, so that incompatible facts (such as the two marked on the bottom edge) simply do not have a greatest lower bound. In the figure at lower left, the line marked a represents atomic facts about some object a, and the shaded region contains complex facts about a. Finally, the figure at lower right is the lattice resulting from removing all facts about a; note that the two marked facts are incompatible in this lattice, because their greatest lower bound in the original lattice occurs in the shaded region under a.

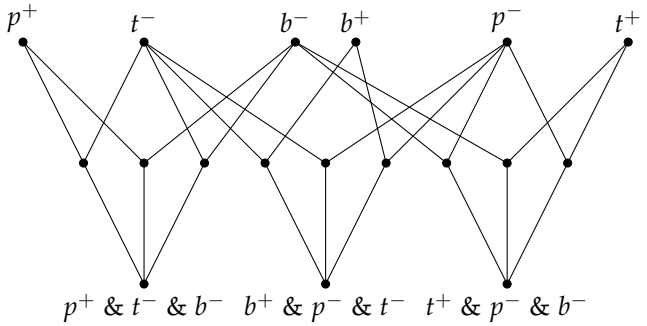


Figure 6.3: Full data model for Walt's situation, in 'partial meet' style.

Walt's pocket (p^+, p^-). I give the picture in terms of the *partial* meet operation &; if f and g have no greatest lower bound in the picture then $f \perp g$.

In this tiny model not all atomic facts are mutually compatible. By the requirements on the interpretation function of a first-order data model, t^+ is incompatible with t^- (this means that it follows directly from Walt's data that $\neg\text{keys}(t)$, as we would hope). But as well there is no fact in the lattice for t^- & b^- & p^-: the structure of the lattice encodes the natural law that the keys must be *somewhere*. The lowest facts in the lattice each represent an entire possible world: any other fact is either included in the world or incompatible with it. That there is no such fact where the keys are nowhere is how the natural law "the keys must be somewhere" is encoded in the model.

This structure is what lets Walt draw non-trivial conclusions from his data: if he observes that the keys are on the table then he knows they must not be anywhere else; if he observes that both the table and his pockets are empty then he knows the keys must be in the basket.

This is not enough to get us Walt's belief-under-assumption, though: if he has only the data t^- (the keys are not on the table) then *must* keys(b) does not hold because in at least one extension of Walt's data set the keys turn out to be in his pocket.

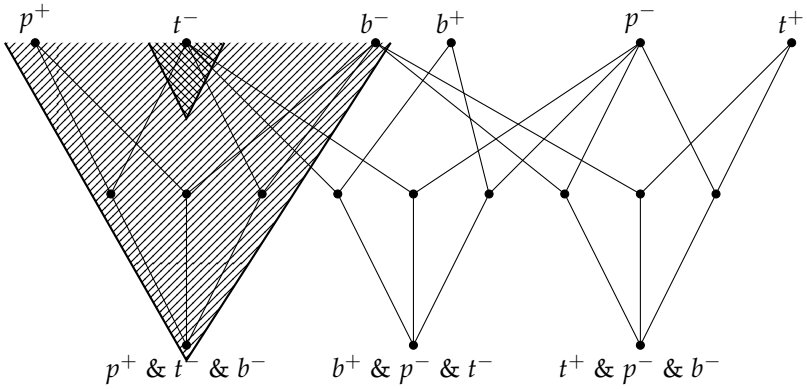


Figure 6.4: Walt's situation, with data. The inner crosshatched region is his present data set (t^- and nothing else); the larger shaded region is a data set extending this, which supports $\neg b$. Because this larger data set exists, his *current* data set does not support *must* b.

Now let us apply the restriction: Walt is aware of only the table and the basket as possible key hiding places, so we remove all facts that are about his pockets. Here is the result:

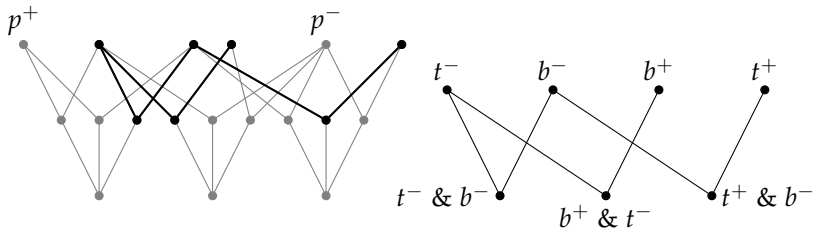


Figure 6.5: Walt's situation under naïve assumptions. The left-hand figure shows shows the original lattice in grey and the facts that are not about p in black (the reader can check that every fact not in black can be traced back up to one of p^+ or p^-). The right-hand figure shows the same restricted lattice, slightly rearranged and labelled.

Walt's data set is still the same singleton $\{t^-\}$. It no longer supports *may* keys(p) (since p^+ is no longer in the data lattice), but it *still* does not support *must* keys(b). The problem now is the possible fact where the keys are neither on the table nor in the basket ($t^- \,\&\, b^-$). In fact Walt seems to hold it possible that the keys are nowhere![4]

The problem seems to be that the structure of a data lattice enforces natural laws (like "The keys are always somewhere") at a global level, with respect to the full domain of objects. A natural law is a *must*-statement that follows from the empty data set: even before Walt opens his eyes in the morning he knows that his keys must be somewhere. In the restricted domain without Walt's pockets, the fact $t^- \,\&\, b^-$ contradicts that natural law; when Walt assumes that only t and b exist, he should not consider this a possible fact at all. But our procedure of removing facts about objects Walt is unaware of takes no notice of natural laws: $t^- \,\&\, b^-$ is a possible fact because it occurs in the (full) data lattice.

Here is another way to see the problem. The fact that is causing problems shows up in the restricted data lattice as the base of a maximal filter: it completely describes the properties of a small world with a two-element domain. But in fact *no such small world exists*, because it would violate the natural law. The fact "The keys are neither on the table nor in the basket" can *only* occur as part of a *larger* world, with a large enough domain (in the original model, the one where the keys are in Walt's pocket). The data lattice has concealed the distinction between a fact that could be a complete description of a (small) world (such as $t^+ \,\&\, b^-$) and one that cannot (such as $t^- \,\&\, b^-$).

[4] I have not yet defined quantification in a restricted model, but the proper definition must use the *smaller* domain if Walt's assumptions of absence are going to be reflected in the quantified sentences his data supports. We will see the formal definition later, once the problem of enforcing natural laws has been solved.

2 · *The proposal*

A possible-worlds semantics with small worlds can represent properly the natural law that the keys are always somewhere, but it cannot distinguish between Walt's immediate data and the conclusions he draws from it (formulae φ and *must* φ). The data semantics model lets us do that, but it cannot combine assumptions of absence (the restriction we tried in the previous section) with natural laws since it does not distinguish which facts are candidate small worlds and which are not.

So we will use *both* frameworks.

The proposal is shown in schematic form in Figure 6.6. We construct two models side-by-side: one is a first-order data model, and one is a possible-worlds model with small worlds. Each encodes the same set of possibilities: each fact in the data lattice is true in some world, and each world corresponds to some particular fact.[5]

Now we will use the strengths of the two models in parallel. Walt's data

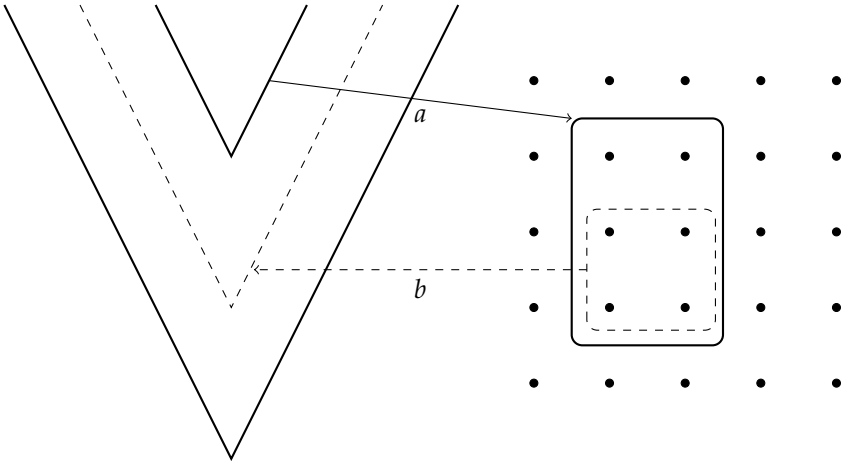


Figure 6.6: The round-trip model. The two halves of the diagram represent a data model (on the left) and a possible worlds model (on the right). Arrow a transforms the agent's data set into an information set in the possible worlds model; arrow b transforms the agent's assumptions (within that set) into a restriction on the data lattice.

[5] Following [Vel81], a world should correspond to a *maximal (proper) filter* on the data lattice, instead of a fact. I take facts instead of filters because they do well enough, so long as the worlds are finite: the fact corresponding to a small finite world is just the conjunction of the atomic facts true in that world, and this generates a filter corresponding to the world. Crucially, though, a small world need not correspond to a *maximal* filter: it omits all facts about objects outside its domain.

lives in the data lattice, but his assumptions will be calculated in the possible worlds model. These assumptions in turn correspond to a smaller data lattice (the more assumptions he holds, the less possibilities he imagines and the more conclusions he draws from his data); the original data interpreted in this smaller data lattice describes Walt's beliefs.

First we need to construct our parallel models. Then reading off Walt's beliefs-under-assumption involves three steps:

1. Data to information: translate Walt's data (in the data lattice) into information (a subset of the possible worlds);
2. Information to assumption: apply an assumption-of-absence on the possible worlds;
3. Assumption to data: translate the resulting assumption into a restriction on the data lattice.

2.1 · *Parallel models*

The small-worlds side of the parallel pair of models is inspired by the object-based unawareness structures of [BC07]. Only "inspired by", though, because just as with the models we began with in Chapter 2, I take a single-agent belief-set view rather than the multiagent relational Kripke semantics.

Definition 6.6: Small worlds model. *A* small worlds model *is a structure* $M = \langle W, \Omega, D, \mathrm{dom}, V \rangle$ *where*

1. W *is a set of worlds;*
2. Ω *is as usual the vocabulary of predicates;*
3. D *is a set of objects, the* domain *of the model (not all objects will exist at every world);*
4. $\mathrm{dom} : W \to \mathscr{P}(D)$ *is the* domain function*, assigning to each world the objects that exist at that world; and*
5. V *is an interpretation function: for each world* w *and predicate* $P \in \Omega$*,* $V(w)(P) \subseteq \mathrm{dom}(w)$ *gives the objects* from the domain of w *that have property* P *at* w*.*

The semantics at each world is given by standard three-valued Kleene definitions, with partiality introduced by the small domains: Px is undefined at w if the variable assignment maps x to an object not in $\mathrm{dom}(w)$. Quantification also works on these domains: $\forall x \varphi$ is true at w if $\varphi[x/a]$ is (defined and) true at w for all $a \in \mathrm{dom}(w)$. (We don't need to be too concerned with these details, because we will only be checking Walt's beliefs about quantified sentences on

the data lattice side of the system.) As in previous chapters, if $B \subseteq W$ is a set of worlds from M then M, B supports (non-modal) φ if all worlds in B support φ, and supports *might* φ if some world $w \in B$ supports φ; we interpret the first as "the agent believes" and the second as "the agent holds possible" (both possibly implicitly).[6]

I *do* need to say explicitly what Walt's awareness of objects involves, and how we can describe his explicit beliefs. Recall from Chapter 2 that $X\varphi$ is to be read as "the agent explicitly believes that φ". The notation I give is more suggestive than comprehensive, however it will suffice for our purposes.[7]

Definition 6.7: Object-based awareness and explicit belief. *Let $\mathfrak{M}$ and M be respectively a first-order data model and a small worlds model that share the vocabulary Ω and domain D. An* awareness state *for the agent is a pair $\sigma = \langle \Xi, D_\sigma \rangle$ with $\Xi \subseteq \Omega$ and $D_\sigma \subseteq D$.*

The agent is aware of a formula φ according to σ if the only non-logical vocabulary occurring in φ consists of predicates in Ξ and objects in D_σ.

If $\mathfrak{d}$ is a data set for $\mathfrak{M}$, then we write $\mathfrak{M}, \mathfrak{d}, \sigma \models X\varphi$ if $\mathfrak{M}, \mathfrak{d} \models \varphi$ and the agent is aware of φ according to σ. If B is a set of worlds from M then we write $M, B, \sigma \models X\varphi$ if M, B supports φ and the agent is aware of φ according to σ.

The predicate awareness is included only for completeness; the interesting part of the system is Walt's awareness of objects. It can be that Walt explicitly believes $\exists x : Px$ (that is, he believes the formula and is aware of it), but he is not aware of any object a of which he holds possible Pa. An example is (2-b) above.

It may be useful to picture the state space of a small worlds model as shown in Figure 6.7 overleaf: as divided into notional 'subspaces', each of which contains worlds sharing the same domain. Walt's assumptions will typically confine his belief set to one such subspace, the one containing only the objects he is aware of.

We can construct a first-order data lattice from a small worlds model, by taking partial interpretation functions as the facts in the lattice. (I will examine what exactly the correspondence between the two models is in a moment.) A partial interpretation function for a first-order language is just the natural generalisation of a partial propositional valuation: it decides for some properties P and objects a whether a has property P, but not necessarily for all pairs. Here

[6] Small worlds models support *might* but not *must*, since *must* φ would be equivalent to φ. Data models support *may* and *must*. The distinction between *might* and *may* is irrelevant, except as far as it serves as a reminder of which kind of model the formula is being evaluated in.

[7] As well as suggestive it is perhaps somewhat misleading: $\mathfrak{M}, \mathfrak{d}, \sigma \models X\varphi$ means not that the data verifies that the agent believes φ but that the agent believes (explicitly) that the data verifies φ. This is why I include no clause for $=\!|$; instead we will use the negation clause from data semantics: $\mathfrak{M}, \mathfrak{d}, \sigma \models X\neg\varphi$.

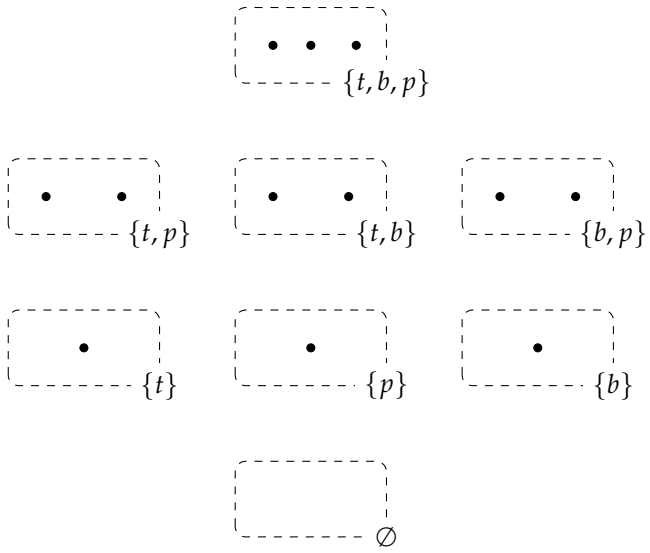


Figure 6.7: Small-worlds model with notional subspaces. Each subspace is labelled with the domain of its worlds. Not every subspace contains worlds (it is impossible for a world with empty domain to satisfy the natural law "the keys are somewhere"). The subspace with domain $\{t, b\}$ contains two worlds: in one the keys are on the table, in the other they are in the basket.

is a definition, heavily simplified from [Mus95].[8]

Definition 6.8: Partial interpretation function. *Let D be a set of objects. A* partial property *is a pair $\langle P^+, P^- \rangle$ where $P^+, P^- \subseteq D$ and $P^+ \cap P^- = \emptyset$. P^+ is the* positive extension *of P (the objects that have the property), P^- the* negative extension *(the objects that do not have the property) and $D \setminus (P^+ \cup P^-)$ its* gap *(objects undefined for P).*

Partial properties come with a natural notion of precision: $P_1 \sqsubseteq P_2$ ("P_1 approximates *P_2") iff $P_1^+ \subseteq P_2^+$ and $P_1^- \subseteq P_2^-$.*

Let Ω be a vocabulary of predicate symbols, and D as before a domain. Then a partial interpretation function *is a function f mapping each element of Ω to a partial property on D: for each $P \in \Omega$, $f(P)^+$ is the positive extension and $f(P)^-$ the negative.*

We lift the approximation order to interpretation functions, relative to a vocabulary Ω: if f and g are partial interpretation functions then $f \sqsubseteq g$ iff for every predicate

[8] I give only definitions for properties; Muskens considers relations in full generality. He also allows over-determined truth values ("both true and false") as well as under-determined ("neither true nor false"); I allow only the latter.

$P \in \Omega$, $f(P) \sqsubseteq g(P)$.

We can think of the valuation $V(w)$ of a small world w as a partial valuation in this sense: one in which for all predicates P, $V(w)(P)^+ \cup V(w)(P)^- = \mathrm{dom}(w)$ (that is, each predicate is decisive for exactly the elements in the small world's domain): the valuation itself gives the positive extension of each predicate, and the negative extension is the rest of the domain of that particular world. In particular, if f is a partial interpretation function and w a world in a small worlds model, then it makes sense under this interpretation to ask whether $f \sqsubseteq V(w)$. The approximation order on interpretation functions has the structure of a meet semi-lattice, which we will use to interpret a small worlds model as a data lattice.

Definition 6.9: Worlds as data. *Let $M = \langle W, \Omega, D, \mathrm{dom}, V\rangle$ be a small worlds model. Then the first-order data model corresponding to M, $\mathrm{data}(M)$, is $\langle \mathfrak{L}, D, \mathcal{I}\rangle$ with $\mathfrak{L} = \langle \mathcal{F} \cup \{\bot\}, \circ, \bot\rangle$ where:*

- *D is taken over unchanged;*
- *$\mathcal{F}$ is given by*

  $$\{f \text{ is a partial interpretation function for } \Omega \text{ on } D \;;\; \exists w \in W : f \sqsubseteq V(w)\};$$

- *$f \circ g$ is the smallest partial valuation[9] $h \in \mathcal{F}$ such that $f \sqsubseteq h$ and $g \sqsubseteq h$, or $\bot$ if none such exists; and*
- *$I(P, +, a)$ is the smallest partial interpretation $f \in \mathcal{F}$ such that $f(P)^+ = \{a\}$, and $I(P, -, a)$ is the smallest partial interpretation g such that $g(P)^- = \{a\}$, or (in both cases) $\bot$ if none such exists.*

The meet operation $\circ$ combines the information in f and g: notionally, for any predicate P, $(f \circ g)(P)^+ = f(P)^+ \cup g(P)^+$ and $(f \circ g)(P)^- = f(P)^- \cup g(P)^-$. The result may be $\bot$ for either of two reasons. One is that f and g may be strictly incompatible: if some object appears in the negative extension of $f(P)$ and the positive extension of $g(P)$, for instance, there is no partial interpretation function that could be their meet. The other reason is that f and g may be *informationally* incompatible, within W: it may be that no world in W contains their joint information. If that is so then although we could construct a partial interpretation function by unifying the two, that function simply does not appear in $\mathcal{F}$, and again the two are incompatible according to the data lattice. (Strict incompatibility is built into our very notion of what a property

[9] This definition may seem surprising since the meet operation selects *greatest lower* bounds, while this constructs *least upper* bounds. However the approximation ordering runs opposite to the information ordering: $f \sqsubseteq g$ iff $g \leq_{\mathfrak{L}} f$, that is, f approximates g iff g includes (the information of) f. The operation $\circ$ is a meet (greatest lower bound) operation on $\leq_{\mathfrak{L}}$, and a join (least upper bound) on $\sqsubseteq$.

is: no object can both have a property P and not have it, at the same time. Informational incompatibility recognises whatever information is contained in the model: if the model contains the natural law "twins have the same eye colour" then the fact that Twin A has blue eyes is informationally incompatible with the fact that Twin B has brown eyes.)

The interpretation function gives us the atomic facts: the fact that a has property P is represented by a partial valuation f that has $f(P)^+ = \{a\}$ but is decided on *nothing else* ($f(P)^- = \emptyset$, and for all other predicates Q, $f(Q) = \langle\emptyset, \emptyset\rangle$). Again, the partial interpretations coming from M may not include one in which a has property P; in that case the atomic fact picked out by the interpretation is the impossible fact $\bot$.

The correspondence I have defined is not any kind of equivalence. Two distinct small-worlds models can produce the same data lattice. (Suppose $V(w) \subseteq V(w')$ and $\mathrm{dom}(w) \subset \mathrm{dom}(w')$. Then the data models constructed from $\{w'\}$ and $\{w, w'\}$ (with the same domains, valuations, &c.) are identical, but the two small worlds models will make different possibility modal formulae true.) There may be formulae true under *must* in the data lattice that are not true everywhere in the small worlds model. (In a small world where a does not exist it does not have any properties, so not even $Pa \vee \neg Pa$ holds.) I will give later on a constraint on the acceptable small worlds models that narrows the gap, but it will never be completely bridged. And indeed this is not surprising: we need the two models exactly because some operations (such as forming assumptions of absence) can be accomplished in the one that cannot be accomplished in the other.

Let us see how these operations are accomplished.

2.2 · *Data to information*

The first thing we have to do is transfer Walt's data to the small worlds model. That's relatively easily accomplished: the worlds he holds possible in the small worlds model are just those that validate his data (and contain all the objects he is aware of).

Definition 6.10: Data set as information. *Let $M = \langle W, \Omega, D, \mathrm{dom}, V\rangle$ be a small worlds model and $\mathrm{data}(M) = \langle \mathfrak{L}, D, \mathcal{I}\rangle$ its corresponding data model, with $\mathfrak{L} = \langle \mathcal{F}, \circ, \bot\rangle$. Let $\mathfrak{d} \subseteq \mathcal{F}$ be the agent's data, a data set for $\mathfrak{L}$, and $\sigma = \langle \Xi, D_\sigma\rangle$ his awareness state ($D_\sigma \subseteq D$ is the objects he is aware of).*

The agent's data set as information in M, written $\mathrm{info}^\sigma_M(\mathfrak{d})$, is given by

$$\mathrm{info}^\sigma_M(\mathfrak{d}) = \{w \in W \,;\, D_\sigma \subseteq \mathrm{dom}(w) \wedge \forall f \in \mathfrak{d} : f \sqsubseteq V(w)\}.$$

There are two parts to the condition, which a world w has to satisfy in order to be part of the information state. The first ($D_\sigma \subseteq \mathrm{dom}(w)$) says that the agent only holds a world possible if it contains all the objects he is aware of (if he

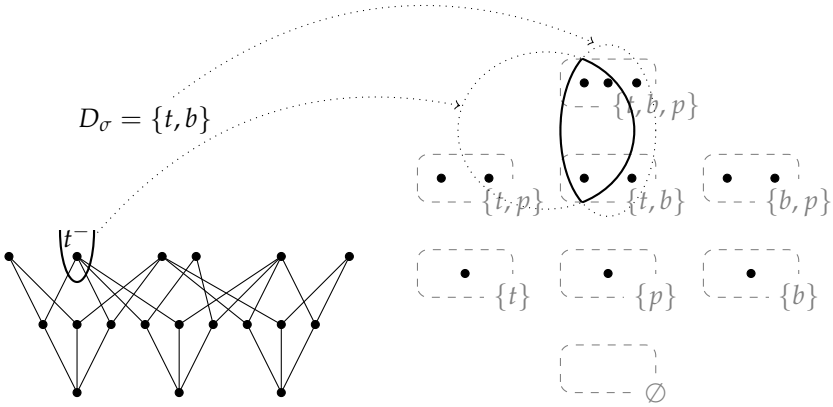


Figure 6.8: Walt's data and awareness as information. His awareness of $\{t, b\}$ means only worlds in two of the subspaces are candidates (the higher of the two arrows). His data picks out some of these (the lower arrow); the excluded worlds in each subspace his data intersects have the keys on the table.

is aware of his pocket, it exists in every world he holds possible). In a sense awareness here *is* information: being aware of an object corresponds to the information that that object exists. The second condition, that all facts in the data approximate the valuation at that world, says that all the data must be satisfied at every world he holds possible.

2.3 · *Information to assumption*

Once we have the information set of small worlds, we can apply the assumptions of absence. The simplest version is just to take only those worlds whose domains are exactly the objects the agent is aware of. That would suffice for example (1) (in all of those worlds, if the keys are not on the table they are in the basket) but not for example (2): there Walt should be forced to realise that there is something he is unaware of. Indeed, Walt's information set once he discovers that both table and basket are empty includes no worlds with two-element domains.

Instead we use DOMAIN CIRCUMSCRIPTION (a notion stemming from early research in non-monotonic reasoning for artificial intelligence [McC80]). From Walt's information set we take the worlds with *smallest* domains: these contain the least number of objects necessary to produce his data.

DEFINITION 6.11: Domain circumscription. *Let* $M = \langle W, \Omega, D, \mathrm{dom}, V \rangle$ *be a small worlds model and* $S \subseteq W$ *a set of worlds. The* SMALLEST WORLDS IN S*, written*

$\min_D(S)$, *are:*

$$\{w \in S \,;\, \neg\exists w' \in S : \mathrm{dom}(w') \subset \mathrm{dom}(w)\}.$$

Walt's assumptions of absence are represented simply by circumscribing his information set: his belief set (under assumptions) is the set of smallest worlds in his information set. If there are any worlds in his information set whose entire domains he is aware of, then these will be the only circumscribed worlds in his belief set: in that case his assumptions of absence say exactly that nothing he is unaware of exists. In cases like (2) above, though, Walt's information is incompatible with this exact assumption; in that case the circumscribed worlds each contain the objects he is aware of, as well as the minimum number of *extra* objects needed to make a 'self-contained' possible small world.

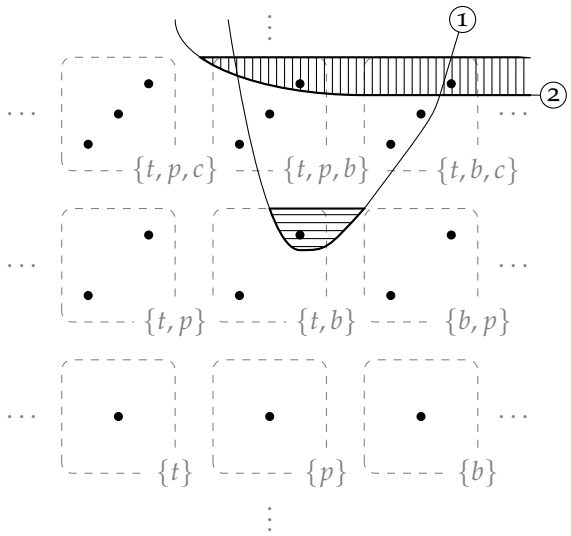


Figure 6.9: Assumptions and subspaces. I have included some extra objects (c is for "couch"), so the figure shows only a part of the whole model. Suppose Walt is aware of only t and b. The line marked "1" gives his information at (1-a): he knows the keys are not on the table but nothing more. The line marked "2" gives his information at (2-b), after learning that they are not in the basket; note that every world he holds possible must contain some object he is unaware of. The horizontal shading gives the result of domain circumscription on the information in 1: Walt concludes that there are no objects he is unaware of. The vertical shading shows the result of domain circumscription on 2: Walt concludes that there is *exactly one* object he is unaware of, namely, the place the keys are hiding.

Assumptions of absence get us partway to representing both examples. If Walt is aware of only the table and the basket, and his data is just that the table is empty, his assumptions will limit him to considering only a single world (in our rather impoverished model, that is): and in that world the keys are in the basket. If he is also aware of his pockets, though, he will have two possible worlds in his assumption set.

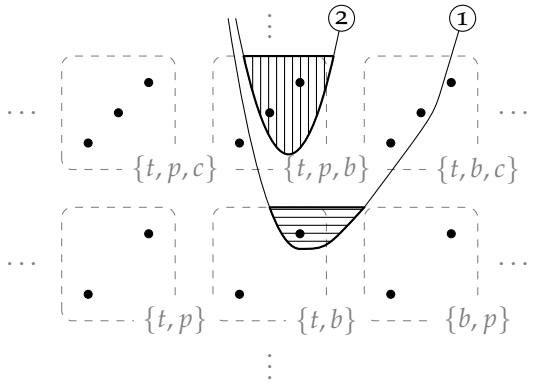


Figure 6.10: Assumptions and awareness. In 1 Walt is aware only of $\{t, b\}$ and knows that the keys are not on the table; the horizontal shading gives his assumption of absence. In 2 he has become aware of his pockets, without gaining any data (as in (1)); vertical shading shows his new assumption of absence.

The second example is more interesting. In that case Walt's information rules out all the two-element small worlds (in no such world are both table and basket empty). Domain circumscription gives us *all* the three-element worlds that support his data. In a properly rich model this would be a large number of worlds, each of which contains one extra object functioning as the hiding place of the keys (down the back of the couch, dropped on the bathroom floor, &c.). Walt is unaware of all these objects, however: he is aware of (and believes) the formula $\exists x : \text{keys}(x)$ but for no object *a that he is aware of* does he hold it possible that $\text{keys}(a)$.

I have kept the running example very small, since otherwise I can't draw pictures to properly represent it. This means that Walt holds an *implicit* belief that the keys are in his pocket, simply because no other possibility exists in the original, full-sized model (that is, the pair of small worlds model and data

model). But this original pair plays the same part that the model of reality or the background model of previous chapters did: it represents what Walt would hold possible 'in the limit' if he became aware of every object that exists. If we are constructing a real model of reality, as opposed to a toy example for the purposes of drawing diagrams, then as well as the three hiding places we have explicitly considered, there exist many many more: down the back of the couch, dropped on the bathroom floor, in the sugar jar or the dishwasher, and so on. If there are enough of such other possibilities, then Walt's *implicit* information will not help him even in the sense that unconscious information guides our day-to-day behaviour. At best, the range of possibilities he unconsciously holds possible will guide him in starting a systematic search.

This is partly because Walt is unaware of whichever particular object holds the keys in each of the worlds he holds possible. But if this is to make any sense, he must also be unaware of the *properties* of these objects that would otherwise serve to identify them (as represented by Ξ, the predicates the agent is aware of). Walt does not explicitly hold possible that the keys are down the back of the couch (if he is unaware of the couch) but he could, in principle, explicitly hold possible that the keys are down the back of some object with the properties of a couch. This doesn't make any psychological sense, so we have to assume that Ξ and D_σ are not entirely independent: thinking about certain properties will call certain objects to mind. We need such a story if we are to (properly; formally) avoid Walt having *explicit* disjunctive information about the properties of the place the keys are hiding: they are either behind something with the properties of a couch or on something with the properties of a bathroom floor or .... I will not give the story in any detail, but it is worth noticing that there are some properties of these objects that Walt *may be* aware of, even though he is unaware of the objects themselves. "Containing the keys" is only represented by a property in our models for the sake of expediency, but "having a cavity large enough to contain the keys" looks much more like the real thing.

If I am permitted this laxity as regards disjunctive information, then we have almost successfully represented our two examples. Walt holds the right beliefs at every step: at (1-a) he believes the keys are in the basket, while after finding they are not at (2-b) he believes they are somewhere else (that he is not aware of). After becoming aware of his pocket he again believes he knows where the keys are (in his pocket), since only those three-element worlds containing table, basket and pocket (the objects he is aware of) make it into his assumption set.

There is still one thing to do though: the small worlds model in which we have achieved this does not represent the difference between direct and indirect conclusion, so we cannot separate Walt's φ from his *must* φ. We want to do so, however, because their stability conditions will be quite different: if we have

done our work properly *must* φ will be unstable under shifts of awareness while plain φ will not. This is in fact not quite the whole story: universally quantified formulae behave differently. To be able to talk about the distinction, though, we need to transfer the assumption set back into the world of data semantics.

2.4 · *Assumptions into data*

In fact we already have almost the definition for doing so: Definition 6.9 showed how we can transform a small worlds model into a data lattice. If we apply the same construction just to the worlds in Walt's assumption set, we will have nearly what we want: a data lattice representing his assumptions. We must be a bit careful about the *direct* effects of unawareness though: the facts of the restricted data model should only be about predicates that Walt is aware of, and the domain should not include any objects that are not in the small worlds.

Definition 6.12: Assumptions as data. *Let $M = \langle W, \Omega, D, \mathrm{dom}, V\rangle$ be a small world model and $S \subseteq W$ any set of worlds. Let $\sigma = \langle \Xi, D_\sigma\rangle$ be an agent's awareness state (with $\Xi \subseteq \Omega$ and $D_\sigma \subseteq D$).*

The first-order data model of S as assumptions is $\langle \mathfrak{L}, D', \mathcal{I}\rangle$ with $\mathfrak{L} = \langle \mathcal{F} \cup \{\bot\}, \circ, \bot\rangle$ where

- *D' is the restricted domain $\bigcup_{w\in S} \mathrm{dom}(w) \cup D_\sigma$ (the union of the domains of worlds in S and the objects the agent is aware of);*
- *$\mathcal{F}$ is given by*

  $$\{f \text{ is a partial interpretation function for } \Xi \text{ on } D \;;\; \exists w \in W : f \sqsubseteq V(w)\};$$

- *$\circ$ is defined as in Definition 6.9; and*
- *$\mathcal{I}$ is likewise defined as in Definition 6.9, but restricted to the predicates in Ξ and the objects in D'.*

Write $\mathrm{data}^\sigma_M(S)$ for the data model formed by this construction.

Note that $\mathrm{data}^\sigma_M(S) = \langle \mathfrak{L}_1, D_1, \mathcal{I}_1\rangle$ is *not* the same structure as $\mathrm{data}(M) = \langle \mathfrak{L}_2, D_2, \mathcal{I}_2\rangle$; viewed in the right way it is a *substructure*. If we view the two data lattices with *partial* meet operations, $\mathfrak{L}_1 = \langle \mathcal{F}_1, \&_1\rangle$ and $\mathfrak{L}_2 = \langle \mathcal{F}_2, \&_2\rangle$, then $\mathcal{F}_1 \subseteq \mathcal{F}_2$ and $f \;\&_1\; g = f \;\&_2\; g$ wherever the former is defined. $D_1 \subseteq D_2$, of course, and $\mathcal{I}_1$ is defined on a subset of the atomic formulae that $\mathcal{I}_2$ is defined on (restricted by the unawareness recorded in σ) but agrees with it on that subset.

Now we are ready to do the whole round trip.

Definition 6.13: Assumptions of absence. *Let M be a small world model and $\mathfrak{M} = \mathfrak{d}(M) = \langle \mathfrak{L}, D, \mathcal{I}\rangle$ its corresponding first-order data model, with $\mathfrak{L} = \langle \mathcal{F}, \circ, \bot\rangle$.*

Let $\mathfrak{d} \subseteq \mathcal{F}$ be the agent's data and $\sigma = \langle \Xi, D_\sigma \rangle$ his awareness state (with $\Xi \subseteq \Omega$ and $D_\sigma \subseteq D$).

Then $\text{info}^\sigma_M(\mathfrak{d})$ *gives his data as information (a subset of the worlds in M),* $\min_D(\text{info}^\sigma_M(\mathfrak{d}))$ *his assumptions under domain circumscription given that information (a smaller subset of the worlds in M), and*

$$\mathfrak{M}^\sigma_\mathfrak{d} = \text{data}^\sigma_M(\min_D(\text{info}^\sigma_M(\mathfrak{d})))$$

represents that set of worlds as a first-order data model ($\mathfrak{M}^\sigma_\mathfrak{d}$ is read as "$\mathfrak{M}$ under awareness given by σ and data given by $\mathfrak{d}$"). Then $\mathfrak{M}^\sigma_\mathfrak{d}, \mathfrak{d}, \sigma$ represents the agent's BELIEFS AND ASSUMPTIONS UNDER UNAWARENESS.

The following figure shows the round trip in detail; in addition, I give names to the intermediate stages that will be more convenient than using the full definitions.

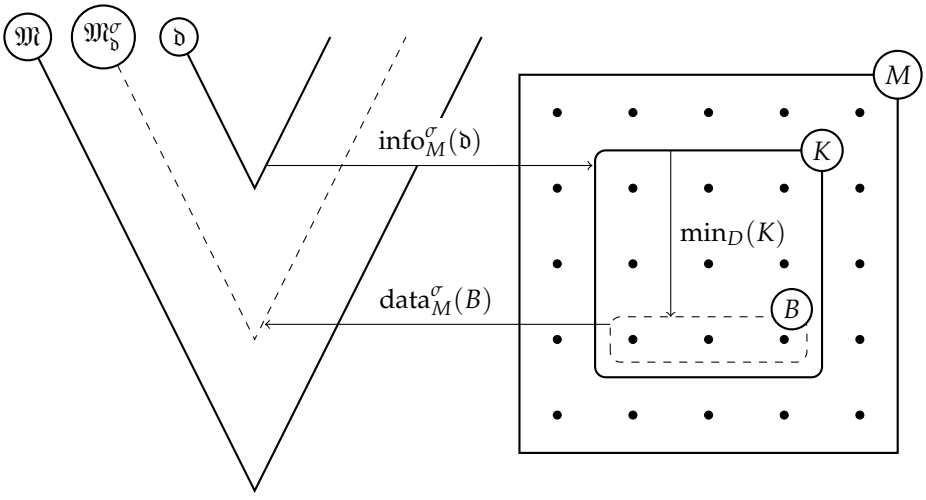


Figure 6.11: The round-trip model, with names for the various components. M and $\mathfrak{M}$ are a matched pair (so $\mathfrak{M} = \text{data}(M)$, not shown) of small-worlds and data models, $\mathfrak{d}$ is the agent's data, and σ (not shown) his awareness state. Then $\text{info}^\sigma_M(\mathfrak{d})$ gives his information set K; $\min_D(K)$ gives his belief set B; and $\text{data}^\sigma_M(B)$ gives his restricted data model $\mathfrak{M}^\sigma_\mathfrak{d}$.

3 · *Properties*

We have the construction; what can we say about it?

3.1 · *Data is preserved*

Firstly, if the agent's awareness is well-behaved then the original data is always preserved by this construction. If $\mathfrak{M}^\sigma_\mathfrak{d} = \langle \mathcal{F}, \circ, \perp \rangle$ then $\mathfrak{d} \subseteq \mathcal{F}$ is guaranteed (modulo a condition on σ, which we will get to in a moment). To see this we have to chain through the definitions.

Let $f \in \mathfrak{d}$ be any fact holding in the agent's data. Obviously f must be a fact in $\mathfrak{M}$, as well. By the construction of the data lattice for $\mathfrak{M}$ (the definition of $\mathfrak{d}(M)$), this means that f holds in at least one world in M. (In particular, since $\mathfrak{M}$ is finite, $\mathfrak{d}$ is generated by a *single* fact; this fact holds in some world of M, which means that at least at one world *all* the facts in $\mathfrak{d}$ hold: $\mathfrak{d}$ is satisfiable in M.)

By the construction of K (the definition of info($\mathfrak{d}$)), our arbitrary fact f holds in all the worlds in K (and K is non-empty since $\mathfrak{d}$ is satisfiable in M). Domain circumscription produces a non-empty set of worlds if the input set is non-empty[10] so $B = \min_D(K)$ is also non-empty, and each world in it satisfies each fact in $\mathfrak{d}$.

The construction of $\mathfrak{M}^\sigma_\mathfrak{d}$ from B should then guarantee that all facts in $\mathfrak{d}$ end up in $\mathfrak{M}^\sigma_\mathfrak{d}$, since they are all satisfied in the set generating it. But does it?

It does not: only those facts *whose predicates the agent is aware of* make it into the data lattice.[11] If the agent's data contains an atomic fact about some property P that he is unaware of, that fact will not survive the round trip.

But how much sense does it make to say the agent has a fact he is unaware of in his data? The agent's data set represents the facts that he has experientially verified; the facts he is directly acquainted with. It seems a reasonable requirement to ask that he be aware of all of them.

Definition 6.14: Awareness consistent with experience. *Let $\mathfrak{L} = \langle \mathcal{F}, \circ, \perp \rangle$ be a first-order data lattice, $\mathfrak{M} = \langle \mathfrak{L}, D, \mathcal{I} \rangle$ a data model, and $\mathfrak{d} \subseteq \mathcal{F}$ a data set for $\mathfrak{L}$. Let $\sigma = \langle \Xi, D_\sigma \rangle$ be an awareness state.*

Then $\mathfrak{M}, \sigma, \mathfrak{d}$ are awareness-consistent *if for every atomic fact $f \in \mathfrak{d}$, there exist $P \in \Xi$ and $a \in D_\sigma$ such that either $\mathcal{I}(P, +, a) = f$ or $\mathcal{I}(P, -, a) = f$. We say also that σ is* consistent with experience.

If Walt's awareness is consistent with his experience, then his data always survives the round trip unscathed. We will assume for the rest of this chapter that the agent's awareness is always consistent with experience. This means that for most non-modal formulae φ, if $\mathfrak{M}, \mathfrak{d} \models \varphi$ then $\mathfrak{M}^\sigma_\mathfrak{d}, \mathfrak{d} \models \varphi$. "Most" here means formulae that get their truth value from individual witness facts in the data set. An atomic formula has this property, as does its negation: *Pa* is

[10] On the assumption that domains are finite.

[11] There is a constraint on objects also, but there at least we are safe: if any fact about a is in the data, then the data is only satisfied at worlds where a exists, so the agent's unawareness of a will not prevent it from participating in the data model under the agent's unawareness.

witnessed by the fact $\mathcal{I}(P, +, a)$ while $\neg Pa$ is witnessed by the fact $\mathcal{I}(P, -, a)$. Complex formulae formed with conjunction and disjunction rely on witnesses if their subformulae do. Existential statements also rely on individual witness facts, but universal statements do not.

Suppose Walt has searched the table, basket and his pockets, and found the keys in his pockets. These are the only three places he is aware of. His data under assumption will directly verify "Everything is either a table, a basket or pockets" ($\forall x : x = t \vee x = b \vee x = p$), while his data interpreted in the original larger lattice will not. This difference is because quantification under assumption goes over the limited domain; some universal statements will be true on this smaller domain that are not true on the larger, on the basis of the same data. (Walt in (1) believes that he is aware of everything, not as a statement with *must* but as a plain non-modal formula.)

Negation, of course, inverts the properties of the quantifiers: the negation of a universal relies on a witness fact and is preserved around the round trip, while the negation of an existential essentially involves the size of the domain and is not so preserved.[12]

If we consider formulae whose only quantifiers are non-negated existentials, then Walt's assumptions cannot make him 'misunderstand' his data. He cannot be mistaken about the facts he has directly witnessed (although he can be wrong about how inclusive his experience was: imagining that he has seen every object that exists, for example).

On the other hand Walt's assumptions *can* cause him to misunderstand what conclusions can be drawn from his data, in much more drastic ways than just believing an unsupported universal. In our running example, Walt believes *must* keys(b) as a conclusion from his data, but it is by no means genuinely certain that the keys are in the basket.

This means that *must* is not T-stable, in the intuitive sense that a formula *must* φ can become false as Walt gains more information (first he believes *must* keys(b), but after checking the basket he no longer does — indeed, he believes $\neg$keys(b)). But the semantics of *must* in these models is given entirely by Veltman's definition, which is T-stable (in the formal, rather than intuitive, sense). So what kind of instability is on display here?

[12] Strictly speaking universals and negated existentials *are* preserved, or at least T-preserved, around the round trip, since if they are verified in $\mathfrak{M}, \mathfrak{d}$ then they are also verified in $\mathfrak{M}^{\sigma}_{\mathfrak{d}}, \mathfrak{d}$. They are not F-preserved or 'undefined-preserved', since they can be *falsified* (or undefined) in $\mathfrak{M}$ and verified in $\mathfrak{M}^{\sigma}_{\mathfrak{d}}$. This is not Veltman's notion of stability, but neither is it the notion considered in the next section; it comes down to the standard observation that universals are not stable under domain extension, since $\mathfrak{M}^{\sigma}_{\mathfrak{d}}$ represents a submodel of $\mathfrak{M}$ with a possibly smaller domain.

3.2 · A kind of update

The answer is that these models combine data semantics proper with a limited form of *update*. Veltman's T-stability refers to possible extensions of a data set within a particular data lattice. But when Walt gains some information, the round trip via the small worlds model and his assumptions changes the data *lattice* against which his data is evaluated.

This is easiest to see when the update involves no new information at all: he just becomes aware of a new object. But it can also hold when the data set grows: some such growth may leave the data lattice unchanged but some will alter it by the effect of assumptions. In (1) Walt gets an information update and then an awareness update; learning that the keys are not on the table leaves his lattice-under-assumption unchanged but increases his data, while becoming aware of his pockets leaves his data unchanged but enlarges the lattice. In (2), on the other hand, learning that the basket *and* the table are empty leads to changes in both data set and data lattice: his data set of course grows to include the facts, and his data lattice grows to include facts holding in three-element worlds.

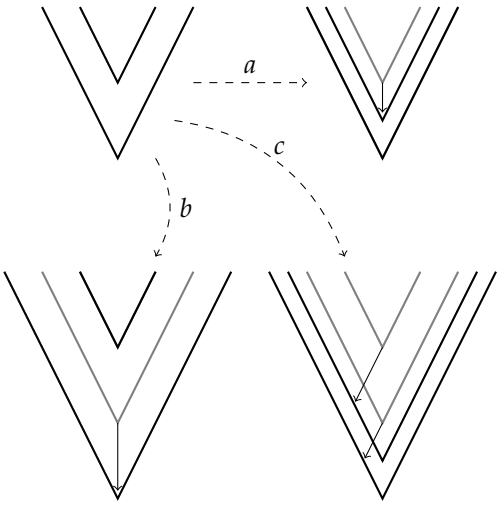


Figure 6.12: Idealised updates. The update a represents pure growth of data; b represents growing awareness (the lattice grows but the data remains the same), while c combines the two.

Veltman's stability refers to the agent's own picture of how his further investigations may proceed; he 'sees' only his lattice-under-assumption, and so a statement with *must* that quantifies universally over this lattice is stable only as

long as the agent 'stays within' that lattice. But the updates can shift him into a *new* lattice, in which the old quantificational facts no longer hold true.

This provides a sort of answer to a comment Veltman makes about his own system.[13]

> [The] truth definitions allow for the background knowledge which one employs to be partial; this is built into the whole idea of successive information states and the like. But the extent to which they do so is limited. Both truth definitions assume in a sense that ones knowledge of the *changes* which ones partial knowledge *could yet undergo* is complete. By freely quantifying over *all* possible extensions of partial information they assume that one, in evaluating conditionals, is in a position to take all of these possibilities into account, that one has a complete knowledge of the structure in question. This is of course not very true to real life. [Vel85, pp. 215–216]

Our agents have complete knowledge of the structure *given their assumptions*, but that knowledge is overturned when their evidence takes them outside what they had assumed to be possible.

3.3 · *Two kinds of stability*

We need a name for the kind of instability that *must* exhibits. Veltman's stability is stability *within* a given data lattice, while the instability of *must* is instability *outside* that lattice. Not just any lattice needs to be considered, though: only those that are in some sense extensions will be important.

Definition 6.15: Data lattice extensions. *Suppose* $\mathfrak{L}_1 = \langle \mathcal{F}_1, \&_1 \rangle$ *and* $\mathfrak{L}_2 = \langle \mathcal{F}_2, \&_2 \rangle$ *are data lattices, with* $\&_1$ *and* $\&_2$ *the* partial *meet operations (that is, we view the lattices as not containing the absurd fact: two facts are inconsistent if their meet is not defined according to* $\&$*). Abusing notation, we write* $\mathfrak{L}_1 \subseteq \mathfrak{L}_2$ *if* $\mathcal{F}_1 \subseteq \mathcal{F}_2$ *and*

$$f \,\&_1\, g = \begin{cases} f \,\&_2\, g & \text{if } f \,\&_2\, g \in \mathcal{F}_1, \\ \textit{undefined} & \textit{otherwise.} \end{cases}$$

We extend the notation to data models *by letting both the lattice and the domain be extended: if* $\mathfrak{M}_1 = \langle \mathfrak{L}_1, D_1, \mathcal{I}_1 \rangle$ *and* $\mathfrak{M}_2 = \langle \mathfrak{L}_2, D_2, \mathcal{I}_2 \rangle$ *are two data models, then* $\mathfrak{M}_1 \subseteq \mathfrak{M}_2$ *iff*

- $\mathfrak{L}_1 \subseteq \mathfrak{L}_2$,
- $D_1 \subseteq D_2$, *and*

[13] In fairness I should point out that Veltman is here concerned with quite a different problem, that of embedding modal and conditional statements. My account has nothing to add here.

- $\mathcal{I}_1 \subseteq \mathcal{I}_2$ *(with the functions also seen as partial: undefined rather than mapping any atomic formula to* $\bot$*).*

Now we can define the two kinds of stability: stability within a lattice (and across data extensions), and stability across lattice extensions with the same data.

Definition 6.16: Intra- and extra-stability. *A formula* φ *is T-intra-stable (F-intra-stable) if for every data model* $\mathfrak{M}$ *and all data sets* $\mathfrak{d}, \mathfrak{d}'$ *for* $\mathfrak{M}$ *such that* $\mathfrak{d} \subseteq \mathfrak{d}'$*, if* $\mathfrak{M}, \mathfrak{d} \models \varphi$ *then* $\mathfrak{M}, \mathfrak{d}' \models \varphi$ *(if* $\mathfrak{M}, \mathfrak{d} \mathrel{=\!\!|} \varphi$ *then* $\mathfrak{M}, \mathfrak{d}' \mathrel{=\!\!|} \varphi$*). This is Veltman's stability notion.*

A formula φ *is T-extra-stable (F-extra-stable) if for every data set* $\mathfrak{d}$ *and pair of data model* $\mathfrak{M}, \mathfrak{M}'$ *such that* $\mathfrak{d}$ *is a data set for both models and* $\mathfrak{M} \subseteq \mathfrak{M}'$*, if* $\mathfrak{M}, \mathfrak{d} \models \varphi$ *then* $\mathfrak{M}', \mathfrak{d} \models \varphi$ *(if* $\mathfrak{M}, \mathfrak{d} \mathrel{=\!\!|} \varphi$ *then* $\mathfrak{M}', \mathfrak{d} \mathrel{=\!\!|} \varphi$*).*

Call a formula existential if it contains no universal quantifiers and any existential quantifiers do not appear under negation. Call it universal if it contains a universal quantifier not under negation. A formula is non-modal if it does not contain *may* or *must*. The following facts about stability are easy to prove:

- Non-modal existential formulae are T-intra- and extra-stable. They rely on witnesses within $\mathfrak{d}$; if the data *or* the lattice gets bigger, those witnesses are still present.
- Non-modal existential formulae are F-intra-stable but not F-extra-stable. Their falsity also relies on having sufficient witnesses within $\mathfrak{d}$, but "sufficient" changes if the expanded data lattice has a larger domain. "There exists a place Walt hasn't looked" is false in a lattice with a small domain (whose elements he has explored) but becomes true based on the same data in a lattice with a larger domain.[14]
- Since $\forall$ and $\exists$ are duals, this means that non-modal universal formulae are F-intra- and extra-stable, and T-intra-stable but not T-extra-stable. Their falsity relies on fixed witnesses within $\mathfrak{d}$; their truth relies on having enough witnesses, so is stable within a lattice but unstable when the domain gets larger.
- *must* φ is T-intra-stable if φ is, but is not F-intra-stable. This is just Veltman's stability notion.
- *must* φ is not T-extra-stable, *even* if φ is both T-intra- and T-extra-stable. For example, let φ be "The keys are in the basket". This is a non-modal

[14] This is the same reason that the round trip does not 'anti-preserve' universal formulae; in that case the domain can shrink, in this case it can grow.

existential, so it is both T-intra- and T-extra-stable. But if Walt's data supports *must* φ but not φ, as in (1), then the new possibilities in a larger lattice may no longer support *must* φ; in this case the natural law "the keys must be somewhere" (implicitly used by Walt in concluding *must* φ) contains a universal quantification over the domain.

- *must* φ is F-extra-stable if φ is F-extra-stable. It relies on a data set $\mathfrak{d}' \supseteq \mathfrak{d}$ falsifying φ; a larger lattice includes this same data set, and if φ is F-extra-stable then in the new data lattice $\mathfrak{d}'$ is still a witness.

Perhaps surprisingly, these stability notions are not all there is to the system I have described. If the only kinds of update we allow are gaining more data and becoming aware of more objects, an update will never make the data set smaller, but it may in fact make the data lattice smaller! There are three possible reasons. One is entirely innocent: gaining data will make the data *lattice* smaller, because the construction only includes facts that are compatible with the data. I will defer discussion of this until Section 3.4.2. Much more worrying is the fact that a pure awareness update can make the data lattice smaller, for either of two reasons. One of these is unintended and must be removed by a further constraint on the models we allow, while the other is an interesting consequence of unawareness.

3.4 · *Stability across updates*

3.4.1 · *Awareness updates only*

Here is a small worlds model that shows the unintended way that an update can make the data lattice smaller.

	$\mathrm{dom}(\cdot)$	$V(\cdot)(P)$
w_1	$\{a\}$	$\{a\}$
w_2	$\{a, b\}$	$\emptyset$

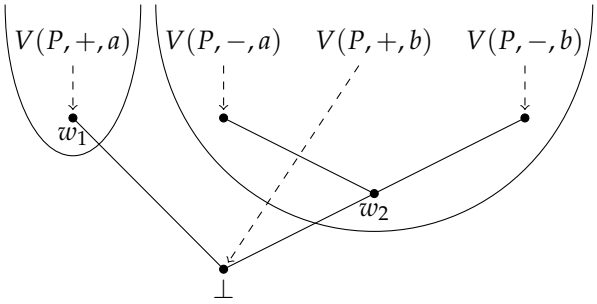


Figure 6.13: Defective small worlds model. Dashed lines give the valuation function, solid lines the information inclusion ordering. The points marked w_1 and w_2 in the data lattice are the facts that 'sum up' those two worlds. The two arcs give the restricted data lattices (in 'partial meet' style) when Walt is aware only of a (the left-hand arc) and aware of both a and b (the right-hand arc).

Now suppose our agent has no data at all, but he is aware of the object a. In that case his assumptions pick out the world w_1 as the only live possibility, giving rise to a data lattice including the possibility that a has property P. If the agent becomes aware of b as well, though, his assumption set shifts to $\{w_2\}$, and within that set there is no possibility that a has property P. His data set (which is empty) is a subset of the new data lattice, but the old data lattice is not.

This means that even formulae that are T-extra-stable are not guaranteed to be preserved by an awareness update. The most familiar type for such a formula is a existential statement with *may* , in our system represented as (for example) *may* $\exists x : P(x)$. While aware only of a the agent *holds it possible* that there is an object which is P, but after becoming aware of b (without learning any new data) he no longer holds this possible.

This doesn't seem to be what we want. And indeed there is something fishy about this particular small worlds model. If the agent learned that $P(a)$, he would also thereby learn that b does not exist. But what if he also became aware of b, which in our system amounts to gaining the information that b *does* exist? This model allows the agent's data to contradict his awareness (as information), while either on its own is satisfiable.

We can avoid this characteristic by demanding that (in effect) the entire domain of the model be real. That is, some world in W has a full domain, and the agent cannot have data that would tell him that some object is not real. This means, in turn, that the 'small worlds' with smaller domains are not *genuine* worlds at all: they are self-contained little parts of the world within which the natural laws can be observed to be functioning. Each can be embedded in a larger world by adding objects, until all the objects in the global domain have been included.

Definition 6.17: Small world embeddings. *We say a small world model* $M = \langle W, \Omega, D, \mathrm{dom}, V\rangle$ ALLOWS EMBEDDINGS *if there is some world* $w \in W$ *such that* $\mathrm{dom}(w) = D$*, and if for each* $w, w' \in W$ *such that* $\mathrm{dom}(w) \subset \mathrm{dom}(w') \subseteq D$*, there is some* $w'' \in W$ *such that* $\mathrm{dom}(w'') = \mathrm{dom}(w')$ *and* $V(w) \sqsubseteq V(w'')$.

What is wrong with the model given in Figure 6.13, according to this definition, is that w_1 is not anywhere 'represented' in the subspace with domain $\{a, b\}$: there is no w'' with $\mathrm{dom}(w'') = \mathrm{dom}(w_2) = \{a, b\}$ where every fact that holds in w_1 also holds ($V(w_1) \sqsubseteq V(w'')$).

If we work only with small world models that allow embeddings, then any existential statement φ that is witnessed at some world with domain $D' \subset D$ will also be witnessed at a world with domain D. Indeed, for each domain D'' such that $D' \subset D'' \subseteq D$, if any worlds at all have domain D'' then one of them

will witness φ.[15]

So is this enough to ensure that may-existentials are stable across awareness updates? Unfortunately still not, but this time for a rather deep and interesting reason.

There is one other case in which the data lattice under assumptions is smaller after an awareness update than it was before, and it is exemplified in example (2). When Walt believes only that the keys must be *somewhere*, without being aware of any place they might be, his assumption set no longer lies in a single 'domain subspace' in the small worlds model. The resulting restricted data lattice includes facts about a huge number of objects (although no single fact is about more than three, not counting the keys themselves). But when he becomes aware of his pockets, all the facts about the couch and the floor of the bathroom and so on are *removed* from the data lattice.

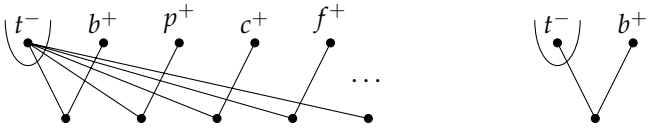


Figure 6.14: Removal of disjunctive imagination under growing awareness. To keep the diagram simple I give a smaller example: in the left-hand lattice Walt is aware only of *one* object (the table t), but has data that the keys are not there (shown as an arc). In the right-hand lattice after becoming aware of one more object, the basket b, all the other possibilities are removed (while his data stays the same).

I don't see any way to avoid this possibility, and I don't really think we want to. The profusion of possible worlds models Walt's complete lack of information —or even assumption— about where the keys are: as far as he can imagine, they could be just about anywhere.[16] The collapse back to a single domain subspace when he becomes aware of his pockets models a kind of persistent optimism: even though his assumptions have proven wrong once, he continues to trust them.

[15] Why quibble about "if any worlds have the domain"? A natural law may prevent certain 'domain subspaces' from containing any worlds at all. Think about the subspace whose domain is just the keys: the natural law "the keys are somewhere" wouldn't be universally satisfied if this subspace contained any worlds. Composite objects are another example: seeing one end of a piece of rope, the agent can be quite sure the other end also exists. That is to say, any world with one end in its domain also has the other, and vice versa; the various domain subspaces that only contain one end are empty of worlds.

[16] It might make sense to combine assumptions of absence with an ordering more like that of previous chapters, but for *objects*: the hiding places behind-the-couch and in-my-other-trousers spring to mind more easily than swallowed-by-the-dog or dragged-into-a-mousehole, say. This would only have an effect in cases like (2).

If we work with models that allow embedding, the extra possibilities that fleetingly become visible are not lost forever. They would again become visible if the agent became aware of the objects they are concerned with (without gaining more data; see the next section). Still it can be that he believes some may-existentials before becoming aware of his pockets but not after; for example, "The keys may be down the back of a piece of furniture." He can always complete such a thought with "Not any of the ones I'm aware of though". We can even imagine (stepping well outside the formal model now) Walt using this kind of disjunctive knowledge of properties to make himself aware of new possibilities: "The keys may be down the back of a piece of furniture…why don't I walk through the house looking for furniture, and check each piece I find?"

3.4.2 · *Stability across data updates*

There is another way in which the data lattice can become smaller after an update, but one which is entirely unproblematic (if a little unexpected). When the agent's data grows, but his assumptions still pick out worlds with the same domains (as in (1) when he checks the table), his restricted data lattice actually *shrinks*.

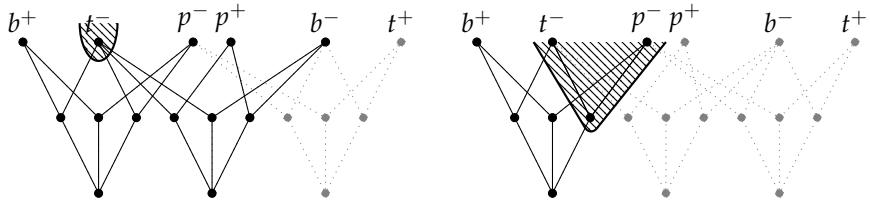


Figure 6.15: How gaining data makes the data lattice smaller. Assume Walt is aware of all three places. On the left, Walt learns that the keys are not on the table; the facts incompatible with this are entirely removed from the data lattice. On the right, he learns furthermore that his pockets are empty; facts incompatible with his new data are likewise removed from the lattice.

What happens is that any facts from the main data lattice that are incompatible with his data do not survive the round trip. If Walt sees that the table contains no keys, the possible fact t^+ will hold in none of the worlds of his information set and so will not appear in the restricted data lattice corresponding to his assumptions.

I've called this "innocent" and "unproblematic" because the presence or absence of t^+ in the data lattice can make no difference to whether $\mathfrak{d}$ supports any formula φ, indirectly (modals) or directly (non-modal formulae). Non-

modal formulae are evaluated with respect to the facts inside $\mathfrak{d}$, but t^+ is not in $\mathfrak{d}$; modal formulae are evaluated with respect to data sets $\mathfrak{d}' \supseteq \mathfrak{d}$, but since such sets must be proper filters and t^+ is incompatible with $\mathfrak{d}$, no such set can ever contain t^+. No kind of formula looks at *alternative* data sets ("If I hadn't learned that the keys are not on the table I would have believed that they may not have been").

Strictly speaking this means that the Veltman-style intra-stability does not tell us much *directly* about our data updates: the data grows but the lattice shrinks, so intra-stability is formally irrelevant. On the other hand, if we call the correct notion "para-stability", then by the above argument a formula is para-stable iff it is intra-stable (T- and F- as required), so intra-stability does in fact capture the kind of stability we want (if indirectly).

4 · *Conclusion*

I have presented a model of *first-order* unawareness, with *assumptions of absence* given by domain circumscription. The fundamental problem of combining data semantics with unawareness, however, is more general: it is the question of the proper representation of natural laws. While I am generally satisfied with the treatment I have given, the extra complications introduced by awareness of objects may obscure some of the general structure of this problem. In particular, the ad hoc and informal arguments I have had to give about implicit disjunctive knowledge of alternatives (in Section 2.3) have to do with the particulars of object-based assumption rather than the general interaction between unawareness and data semantics models; they indicate a difficulty with this particular model but are nothing but a distraction from the most interesting problem. In the rest of this chapter I will say a few more general words about this problem, without specifically relating it to object-based models.

The structure of the round-trip model essentially comes down to solving the problem of representing natural laws under possible unawareness. In data semantics a natural law (or set of laws) is represented directly as a data lattice: facts in the lattice are in accord with the laws, while (pseudo-)facts not in the lattice violate them. (We can talk about a "fact that is not in the lattice" only under the partial meet view: if f & g is undefined then it is a 'fact', or pseudo-fact if you prefer, that is not in the lattice.) The problem in a nutshell is that we need to represent 'the same' natural law applied over different data lattices, generated from different sets of atomic facts.

If a natural law simply *is* a data lattice, then this is strictly impossible: different data lattice, different natural law. But the way we normally think about natural laws, the same law can apply in different circumstances: a natural law is something that *generates* a pattern of presence and absence in a data lattice.

What unawareness shows, I think, is that this generation is not simply a matter of asking which complex facts violate the natural law in question; unawareness shows that a fact cannot in itself violate a natural law. It can only do so within the context of a particular restricted set of possible facts (we might say, in the context of a set of assumptions). This is because the rules we talk about as 'natural laws' themselves come hedged with normality assumptions: everybody has a mother (except those artificial babies they'll be developing over at GenTech in 2050), the keys are always somewhere (unless they've been ground into powder and the powder dissolved in acid), copper sulphate burns with a blue-green flame (unless it's receding close to the speed of light, in which case the colour shifts towards the red end of the spectrum).

The complex fact "I burnt the stuff, and it made a red flame, and it was pure copper sulphate" contradicts the statement "If the stuff is pure copper sulphate, it burns with a blue-green flame." But in the context of unawareness we should not call that statement a natural law: it is a formula whose extension is equivalent to the relevant natural law, on a restricted set of worlds corresponding to assumptions of normality (in particular, that the sample is not receding at high relativistic velocity). The real natural law is an impossibly complex interrelationship between chemistry, physics and biology (the chemistry of burning, the physics of the emission and propagation of light, and the biology of vision), that we are probably not capable of grasping in its entirety and that we do not need to grasp in order to fruitfully use approximations of the law in our day-to-day lives. When we talk about the natural law that copper sulphate burns blue-green, we mean the 'approximate law' that *under assumptions of normality*, copper sulphate burns blue-green.

This is problematic for data semantics, because the fact "The copper sulphate burned with a red flame" doesn't say anything about normal or abnormal circumstances. It violates the approximate law if the circumstances are normal, it does not if they are abnormal (it violates the actual law if the sample is at rest relative to the observer —and if a host of other normality conditions are met— otherwise not). The fact *itself* is compatible with the law, but it leads to the conclusion that conditions must be abnormal. Under unawareness, though, the data lattice might not include the particular abnormalities that would make this fact possible.

In other words, this particular fact should be included in a data lattice that also includes a particular abnormality condition (the possibility of extreme relativistic speed) but not in one that doesn't. Likewise, the fact that both the table and the basket are empty should be included in a data lattice that includes somewhere else for the keys to be, but not in one that doesn't. Each natural law says which *maximal* (conjunctive) facts it allows as possible, not which facts

simpliciter.[17]

The round trip model is an instantiation of this general idea: a natural law is given as a set of *worlds*, each of which stands for a maximal (conjunctive) fact. A small world implicitly represents the normality assumption "and everything else is normal and irrelevant"; when the agent becomes aware of some object a not in a small world w, w no longer counts as a maximal conjunctive fact because facts about a can be added to it (and thus w no longer features in the agent's information set). The embedding constraint comes down to the following: if some fact f is possible (according to the natural laws) under *implicitly* normal conditions (in a small world), it should also be possible under *explicitly* normal conditions (in some larger world).

It must remain a problem for further work to establish whether this general schema can be formalised in some more natural extension of data semantics. Since the general formulation takes in the kinds of phenomena that I have dealt with in the first part of this dissertation, my hunch is that the principal stumbling block will be finding an appropriate representation for assumptions, since domain circumscription is only appropriate for the narrow class of assumptions I have dealt with in this chapter.

[17] Geometric intuitions about partiality orderings can be confusing. A maximal conjunctive fact is a minimal element of the $\leq$ ordering: it lies as far as possible from the atomic facts.

Chapter 7

Conclusions and further work

> It is the nature of an hypothesis, when once a man has conceived it, that it assimilates everything to itself, as proper nourishment; and, from the first moment of your begetting it, it generally grows the stronger by everything you see, hear, read, or understand. This is of great use.
>
> Laurence Sterne, *Tristram Shandy*

If your experience of reading this dissertation has been anything like my experience of writing it, having made your way this far through the various notions and notations you may feel you have rather lost track of the big picture. In this final chapter I will sum up the main ideas that I have attempted (with varying degrees of success) to flesh out as formal systems. Then, in the true spirit of awareness, I will end this dissertation by raising some possibilities that I have not had time to address explicitly, in the hopes of making the reader aware of something he or she has not hitherto entertained.

1 · *Some explicit beliefs*

Awareness is just now making the transition from a new and problematic concept in epistemic logic to a standardised notion that can be applied as a tool in other fields, driven largely by research in the formal economics community. I hope this dissertation contributes to this movement, by showing the potential applicability to problems in formal semantics and pragmatics.

Intuitions about unawareness or attention to possibilities are common in linguistics, but tend to be relegated to the proverbial 'wastepaper basket' of pragmatics. (This is not, I hasten to add, an opinion of pragmatics which I share.) In this dissertation I have tried to formalise some of these intuitions, and to suggest where some existing problems can fruitfully be recast in terms of awareness.

In adapting models from the economics literature for formal semantics and pragmatics I have been concerned with a problem that has received little attention in that community: the influence of unawareness on belief. My notion of ASSUMPTION is the key to representing this influence: assumptions are beliefs which are *only* held because of unawareness, and we can examine the 'assumptive component' of any belief. An interesting feature is that explicit

beliefs, too, typically rely on assumptions (the notion does not reduce to the distinction between implicit and explicit belief).

On a behaviouristic view, assumptions are only distinguishable from 'ordinary' explicit beliefs when they get overturned: drawing attention to an assumption allows the agent to question it, while drawing attention to a belief she already holds explicitly doesn't change her epistemic state. I have adapted update theories from formal semantics to represent this effect, but a lot of work remains to be done here. In particular, all the models I have described are flat single-agent representations; it remains to be seen if these techniques can be easily extended to multi-agent relational systems.[1]

Besides the notion of assumption, the most important notion for linguistics that comes 'for free' with models of unawareness is the context change that comes with attending to new possibilities. This is not the outcome of reasoning, it is not dependent on assumptions of speaker competence, trustworthiness or cooperation, and it cannot be rejected, cancelled or denied. I have applied this notion to some reasonably simple cases (everyday use of *might* and counterfactuals); in the rest of this chapter I would like to suggest some more speculative applications.

2 · *Some unexamined possibilities*

Implicit notions of awareness can be found throughout the formal semantics and pragmatics literature. It is tempting to say that the field has always had *assumptions* about awareness, and it is only now, with growing awareness of awareness itself as a formal notion, that these are becoming ratified as beliefs (or in some cases overturned).

In particular, I believe that some phenomena currently described as accommodation would be better recategorised as effects of dynamic awareness and assumption. Lewis discusses a number of cases which he unifies under the term 'accommodation' in his seminal paper on the subject [Lew79]. His proposed rule of accommodation (the general framework all his examples fall under) is

> If at time t something is said that requires component s_n of the conversational score to have a value in the range r if what is said is to be true, or otherwise acceptable; and if s_n does not have a value in the range r just before t; and if such-and-such further conditions hold; then at t the score-component s_n takes some value in the range r. Page 347

This rule is carefully phrased so as not to make any statement about reasoning,

[1] Some models of awareness dynamics already exist in the economics literature but they are focused on game-theoretic applications and make no allowance for assumptions. [DF09], which I became aware of only very late in the writing of this dissertation, gives a very promising approach somewhat akin to public announcement logics.

but the term 'accommodation' is nowadays used in a somewhat narrower sense. We expect speakers to 'follow the rules'; if their behaviour can only be interpreted as following the rules by changing the score, then we change the score.[2]

Changes in awareness can provide behaviour that matches the technicalities of Lewis's rule, but for reasons that have nothing to do with pragmatic reasoning. Consider the example Lewis gives of relative modality (we have already seen this example in Chapter 4):

> Suppose I am talking with some elected official about the ways he might deal with an embarrassment. So far, we have been ignoring those possibilities that would be political suicide for him. He says: "You see, I must either destroy the evidence or else claim that I did it to stop Communism. What else can I do?" I rudely reply: "There is one other possibility — you can put the public interest first for once!" That would be false if the boundary between relevant and ignored possibilities remained stationary. But it is not false in its context, for hitherto ignored possibilities come into consideration and make it true. And the boundary, once shifted outward, stays shifted. If he protests "I can't do that", he is mistaken. Pages 354–355

This is an instance of accommodation in the extremely abstract formulation given by Lewis: his utterance ("You can put the public interest first") is only acceptable if the possibility it mentions is 'on the table' at least in the minimal sense of being *entertained* by both parties, and if it was not entertained before the utterance then it will be afterwards. This last conditional is vacuous, as the possibility will be entertained after the utterance regardless of whether it was before or not; Lewis's rule does not explicitly come out and say *why* the score changes. But according to the intuitive (modern) notion of accommodation, it should change because the speaker's behaviour is only acceptable if it does; this is not the case, on the awareness account, for this update.

We can see a hint of this in the fact that the 'rule of accommodation' only works one way in this example. If the possibility had already been entertained and excluded ("I have chosen not to put the public interest first. So you see, I must either…") then the rule of accommodation would suggest that Lewis can still make his protest, which he certainly cannot do. Of course there is the hedge

[2] If accommodation is pragmatically driven in this way, then we should be reluctant to accommodate for speakers who we know to consistently break the rules. Edward Gorey's delightful picture-book "The Object-Lesson" begins with the line "It was already Thursday, but his lordship's artificial limb could not be found." Neither Lordship nor limb seem to feature in the rest of the story (if "story" it can be called). My favourite line, "It now became apparent (despite the lack of library paste) that something had happened to the vicar", appears near the end; by that time any reader, no matter how conscientiously cooperative, has long since given up accommodating the various presuppositions that are being invoked with such inventive abandon.

that "such-and-such further conditions" must hold, but we can see where this is going: the correct statement of these conditions for this case will very likely be that the possibility being mentioned may not have been previously entertained. In that case, the rule of accommodation is a completely superfluous (although perfectly correct) description of the phenomena.

There seem to be a number of other examples following the same pattern, so that we can pick out a subclass of Lewis's accommodation phenomena that owe their existence to (something like) awareness (and that are not accommodation in the modern sense: they are not driven by any kind of pragmatic reasoning). What singles out these particular cases is that they have a *preferred direction*, and that direction is towards *more inclusive sets* (of whatever is being quantified over).

There are two cases treated explicitly in these terms in [Lew79]: the section on relative modality already mentioned, and the analysis of vagueness. I have dealt with some of the observations about relative modality already, but I have focused on commonsense uses of *might* and *must*; Lewis applies the accommodation account also to sceptical argument. In general schema the problems of scepticism and vagueness both seem to match rather well to unawareness accounts, but there are also indications that the details may prove problematic. I have been able to formulate some of the questions that should be asked, but not to give them any answers.

2.1 · *Epistemology and the sceptic*

The problem of the sceptic is a clash between two fairly basic intuitions about knowledge: firstly that we have a lot of commonsense knowledge (I know what my name is, where I live, and so on), and secondly that a proposition is not truly known unless all possibility of doubt has been eliminated. The sceptic's argument puts these two intuitions into conflict by raising possibilities that cannot be eliminated, but that appear to undercut our commonsense knowledge; for example, that I have been recently hypnotised and my beliefs about my name and address are in fact incorrect. The possibility cannot be eliminated because, the sceptic argues, it is conceivable that the hypnotism was so perfectly conducted that I cannot tell the difference between my hypnotically induced memories and my true ones.

We have seen already some approaches to representing knowledge under unawareness; in particular the model of [HMSo6] explicitly sets out to represent the limits that unawareness places on potential for knowledge. This approach is vulnerable to sceptical argument, however: if the agent is unaware of the possibility that he is hallucinating, and if under a hallucination nothing he sees is real, then he cannot know *anything* about what he sees. The problem is avoided by not including sceptical possibilities in the models, which is reasonable enough when they are intended for economic applications but is

hardly defensible on epistemological grounds.

An opposite extreme is to take the agent's knowledge as generated from within his assumptions; this would acknowledge that we can know mundane facts like where we live, so long as we do not entertain (or wonder about) sceptical possibilities. Taken without reservation, this principle relaxes all normative standards: it leads to non-factive knowledge (since the agent's assumptions may rule out the actual world), and it implies what Catherine Elgin called the "epistemic efficacy of stupidity", that "stupid people may be in a better position to know than smart ones" [Elg88].

[Lew96] takes an intermediate road between these two extremes: "*S* knows that *P* iff *S*'s evidence eliminates every possibility in which not-*P*—Psst!—except for those possibilities that we are properly ignoring." This definition takes our two problems into account in the "*sotto voce* proviso": "Ignoring" allows assumptions to influence knowledge, while "properly" maintains a normative standard. Stupidity does not grant knowledge because some possibilities ignored by the stupid will not be *properly* ignored (and if the actual world is never properly ignored then knowledge is again factive), but sceptical possibilities do not destroy knowledge unless they are attended to (so when ignoring them I can still know what my address is and similar mundane facts).

If Lewis can be trusted to spell out the normative conditions of "properly", it seems that an unawareness model will be perfect for representing "ignoring".[3] However there are two complications: the first is the source of normative judgements, and the second (more fundamental) is a problem with shifting the meanings of expressions.

2.1.1 · *Multi-agent epistemology*

As soon as we start thinking about knowledge *attributions* the question arises, whose awareness matters? If I am embroiled in a sceptical discussion and, looking out the window, say "Those people out there think they know their own names, but they're wrong: the possibility that they might be hypnotised victims of a hoax undercuts their knowledge", am I wrong?

Lewis wrote "that *we* are properly ignoring", but it is easy to manufacture cases where attributor and attributee attend to different sets of possibilities, and it is by no means clear that a single rule will always provide a sensible answer here. The multi-agent properties of awareness play their part here also: if I do not imagine the possibility that *p* then I certainly do not imagine that *you* imagine the possibility that *p*, so my own assumptions will have an effect on the knowledge attributions I am willing to assert (quite distinct from the question of how these assumptions influence which of those assertions are *true*).

[3] I have contributed a very simple model along these lines to a paper giving a range of analyses of "knowing whether" constructions: [AEJ].

While these questions are interesting and difficult, I presume that some kind of answer is in principle possible (even if it be "it depends"). The next problem, however, indicates the very boundary of applicability of the approach that I have followed throughout this dissertation.

2.1.2 · *Meaning shift under changes in awareness*

All the models I have presented contain a syntactic element, representing some agent's conceptual vocabulary or language of self-ascription of beliefs.[4] The agent comes pre-equipped with a potential vocabulary of concepts, with a logical specification of how those concepts apply across the space of possible worlds. The extreme possibilities introduced by sceptical argument, though, call such a notion into question.

If we are to follow Lewis's schema for knowledge, we need to be able to say when an agent's evidence eliminates a possibility. But the very terms that we use to describe the *evidence* are themselves subject to sceptical attack. If I say (along with Lewis [Lew79, p. 355]), "I know the cat is in the carton — there he is before my eyes," you can reply that my evidence is nothing but the pattern of light arriving on my retina (which might in fact come from an ultra-high-tech projector system, or a deceiving demon); when I concede this you go further: my evidence is in fact nothing more than my sensations (compatible with light patterns on my retina, but also with artificial stimulation of my neurons), and so on. The sceptic's game is always to point out how the commonsense concepts involved even in the very notion of evidence itself rely on unstated assumptions for their effectiveness.

Where this causes difficulties in our formal models is in the representation of the agent's evidence. For instance, if you try to convince me that this is not a hand by raising the possibility that I am hallucinating all my experiences, the atomic formula that I have been interpreting as "My hand hurts" will have to be amended to "I have pain sensations of such-and-such a character", and so on. (Otherwise we can say, with Moore, "I know I have a hand, therefore objects exist".) This behaviour is quite different to the paradigm awareness cases, in which the interpretation of *utterances* may shift over time (as the assumptions giving rise to them become visible) but the interpretation of *sentences* (against the abstract background 'model of reality') remains fixed.

This is not really a new observation, dressed up though it is in the new language of unawareness. Wittgenstein wrestled with the same problem in *On Certainty*:

> That is to say, the *questions* that we raise and our *doubts* depend on the fact that some propositions are exempt from doubt, as it were

[4] The model of [HMSo6] is defined without such a language, but these remarks apply equally there once it is equipped with a logical language interpreted at states in the model.

> like hinges on which those turn. [Wit69, §341]

At first sight it seems like unawareness should apply nicely to the problem, but then the question arises: what should it mean if we draw attention to the hinges?

2.2 · *Vagueness*

The second application of unawareness that I did not manage to spell out is to vagueness and standards of precision. Again, Lewis describes the data as an accommodation effect, and again the resistance to reversing the direction of change seems to indicate something like a growing awareness of a space of possibility:

> One way to change the standards is to say something that would be unacceptable if the standards remained unchanged. If you say "Italy is boot-shaped" and get away with it, low standards are required and the standards fall if need be; thereafter "France is hexagonal" is true enough. But if you deny that Italy is boot-shaped, pointing out the differences, what you have said implies high standards under which "France is hexagonal" is far from true enough.
>
> I take it that the rule of accommodation can go both ways. But for some reason raising of standards goes more smoothly than lowering. If the standards have been high, and something is said that is true enough only under the lowered standards, and nobody objects, then indeed the standards are shifted down. But what is said, although true enough under the lowered standards, may still seem imperfectly acceptable. Raising of standards, on the other hand, manages to seem commendable even when we know that it interferes with our conversational purpose.

He goes on to cite Peter Unger's argument that hardly anything is flat: since "flat" is an absolute term, if a is flatter than b then b is not flat; but for just about anything we can find something flatter, so pretty much everything must not be flat.

The answer that awareness suggests for this conundrum is by no means novel. It might have the virtue, though, provided by every unifying framework: if there is anything to the comparison then we will find that these aspects of vagueness are 'like' other awareness phenomena in more than superficial ways.

2.2.1 · *Standards of precision*

It seems clear enough to me that two different mechanisms are involved in raising and lowering the standards of precision: lowering may indeed proceed by accommodation, but raising goes by something analogous to an awareness

update. It is no coincidence that in objecting that Italy is not boot-shaped you must point out the differences: in much the same way one cannot overturn knowledge by a 'lazy sceptic's argument' "You don't know that because *you might be wrong*" but must draw attention to a particular, specific possibility.

The analogy is however only partial, since whatever it is that these agents should be aware and unaware of, they are not possibilities, concepts, or objects in first-order models, in the sense that I have used these terms. In the case of country borders they might be abstract geometrical figures: a low standard of precision contains a few shapes such as a hexagon, a square, and a boot, while a high standard of precision contains in addition more of the possible finely-varied outlines. In that case becoming aware of new possible shapes raises the standards of precision.

A similar case has been made by Manfred Krifka [Kri07] regarding the inherent vagueness (or implied precision) of numerical terms. Elwood Blues announces "It's a hundred and six miles to Chicago," implying an accuracy of one mile; the same distance could as acceptably be described as 100 miles (under lower standards of precision), while 100 miles can never acceptably be described as 106. Here it is different cognitively salient scales of measurement that the agents must be aware of: the scales with intervals of one mile, five miles, ten miles, fifty miles, and so on are all possibilities. Mentioning 100 miles need not call attention to the one-mile scale, while mentioning 106 miles necessarily does (since the value does not lie on any other scale).

I see two important open questions about using awareness for these sorts of problems. The first is whether we can make a sensible semantics using truth by approximation; "Italy is boot-shaped" would be true, for example, if the actual shape of Italy is *most similar* to the boot out of all the shapes under consideration. Whether this will directly cause problems I'm not sure (I will argue in a moment that the most obvious negative consequence, the sorites paradox, can be ducked using assumptions); it would certainly require reinventing rather a lot of wheels though. The second question is how the combinatorics works. If I argue that France looks more like a star than a hexagon, what effect does that have on the judgement that Italy is boot-shaped? What about if I point out that France's borders are not straight? Things are easier in the numerical case, since we have natural scales we can talk about (mentioning 106 calls attention to *all* distances on the one-mile-unit scale) but of course we do not have such clear-cut and well-defined standards for every vague predicate.

2.2.2 · *The sorites paradox*

Besides standards of precision, the second well-known difficulty with vague predicates is of course the sorites paradox. There too awareness models may have something to offer.

If an account something like the above is correct, then using a particular

standard of precision for some vague predicate P amounts to being aware of only a few of the many possible degrees of P-ness. What could be more natural, then, but that assumptions rule out the intervening degrees?

That is, if Carrie Fisher's character in The Blues Brothers attends only to the ten-mile scale, then she assumes that there is no such distance as 106 miles! This is why we need an approximate (similarity-based) semantics: the actual world (or in this case the actual distance) no longer appears at all in Fisher's set of possibilities.

This certainly stops the sorites argument dead in its tracks. In fact, the very idea of a standard of precision associated with a degree scale may be that the increments on that scale *are* suitably sized to distinguish whether something is P or not-P. (If you are using the one-hair scale then you thereby admit that for some element n on that scale, n hairs is bald while $n + 1$ hairs is not bald.) Of course it raises all sorts of other problems instead...

One has to stop somewhere, however.

> For small erections may be finished by their first architects; grand ones, true ones, ever leave the copestone to posterity. God keep me from ever completing anything. This whole book is but a draught—nay, but the draught of a draught. Oh, Time, Strength, Cash, and Patience! Herman Melville, *Moby-Dick*

Bibliography

[AEJ] Maria Aloni, Paul Égré, and Tikitu de Jager. "Knowing whether A or B". Forthcoming in *Synthese*.

[AGM85] C. E. Alchourròn, Peter Gärdenfors, and D. Makinson. "On the logic of theory change: Partial meet contraction and revision functions". In: *Journal of Symbolic Logic* 50 (1985), pp. 510–530.

[Aum76] Robert J. Aumann. "Agreeing to disagree". In: *The Annals of Statistics* 4.6 (Nov. 1976), pp. 1236–1239.

[BC07] Oliver Board and Kim-Sau Chung. "Object-based unawareness". In: *Logic and the Foundations of Game and Decision Theory, Proceedings of the Seventh Conference*. Ed. by G. Bonanno, W. van der Hoek, and M. Woolridge. 2007.

[BCS09] Oliver Board, Kim-Sau Chung, and Burkhard C. Schipper. *Two models of unawareness: Comparing the object-based and the subjective-state-space approaches*. Working paper 09-3. Department of Economics, University of California, Davis, 2009.

[BJR05] Anton Benz, Gerhard Jäger, and Robert van Rooij, eds. *Game Theory and Pragmatics*. Palgrave Studies in Pragmatics, Language and Cognition. 2005.

[BR07] Anton Benz and Robert van Rooij. "Optimal assertions, and what they implicate. A uniform game theoretic approach". In: *Topoi* 26 (2007), pp. 63–78. DOI: 10.1007/s11245-006-9007-3.

[DF09] Hans van Ditmarsch and Tim French. "Awareness and forgetting of facts and agents". In: *Proceeedings of the 2009 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies (WI-IAT 2009)*. To appear. 2009. URL: http://www.cs.otago.ac.nz/staffpriv/hans/pubs/wliamas_vDF.pdf.

[DLR98] Eddie Dekel, Barton L. Lipman, and Aldo Rustichini. "Standard state-space models preclude unawareness". In: *Econometrica* 66.1 (1998), pp. 159–173.

[Elg88] Catherine Z. Elgin. "The epistemic efficacy of stupidity". In: *Synthese* 74 (1988), pp. 297–311.

[FH88] Ronald Fagin and Joseph Y. Halpern. "Belief, awareness and limited reasoning". In: *Artificial Intelligence* 34 (1988), pp. 39–76.

[FJ07] Michael Franke and Tikitu de Jager. "The relevance of awareness". In: *Proceedings of the Sixteenth Amsterdam Colloquium*. Ed. by Maria Aloni, Paul Dekker, and Floris Roelofsen. 2007, pp. 91–96.

[FJ08] Michael Franke and Tikitu de Jager. *Now that you mention it: Awareness dynamics in discourse and decisions*. ILLC Prepublication ILLC-2008-47. Universiteit van Amsterdam, 2008. URL: http://dare.uva.nl/en/record/285542.

[Fin01] Kai von Fintel. "Counterfactuals in a dynamic context". In: *Ken Hale: A Life in Language*. Ed. by Michael Kenstowicz. MIT Press, 2001, pp. 123–152.

[Fra09] Michael Franke. "Signal to Act: Game Theory in Pragmatics". PhD thesis. Universiteit van Amsterdam, 2009.

[GJS84] Jeroen Groenendijk, Theo Janssen, and Martin Stokhof, eds. *Truth, Interpretation And Information: Selected Papers from the Third Amsterdam Colloquium*. Foris Publications, 1984.

[Gär82] Peter Gärdenfors. "Imaging and conditionalization". In: *The Journal of Philosophy* 79.12 (Dec. 1982), pp. 747–760. URL: http://www.jstor.org/stable/2026039.

[Gaz79] Gerald Gazdar. *Pragmatics*. London: Academic Press, 1979.

[Gil07] Anthony S. Gillies. "Counterfactual scorekeeping". In: *Linguistics and Philosophy* 30 (2007), pp. 329–360. DOI: 10.1007/s10988-007-9018-6.

[Gri67] H. P. Grice. "Logic and conversation". The William James Lectures, delivered at Harvard University. Republished with revisions in [Gri89]; page numbers refer to this edition. 1967.

[Gri89] H. P. Grice. *Studies in the Way of Words*. Cambridge, Massachusetts: Harvard University Press, 1989.

[HMS06] Aviad Heifetz, Martin Meier, and Burkhard C. Schipper. "Interactive unawareness". In: *Journal of Economic Theory* 130 (2006), pp. 78–94.

[HMS08] Aviad Heifetz, Martin Meier, and Burkhard C. Schipper. "A canonical model for interactive unawareness". In: *Games and Economic Behavior* 62 (2008), pp. 304–324.

[HR08] Joseph Y. Halpern and Leandro Chaves Rêgo. "Interactive unawareness revisited". In: *Games and Economic Behavior* 62 (2008), pp. 232–262. DOI: 10.1016/j.geb.2007.01.012.

[Hir85] Julia Bell Hirschberg. "A Theory of Scalar Implicatures". PhD thesis. University of Pennsylvania, 1985.

[Hor72] Laurence R. Horn. "The semantics of logical operators in English". PhD thesis. Yale University, 1972.

[Hor84] Laurence R. Horn. "Towards a new taxonomy of pragmatic inference: Q-based and R-based implicature". In: *Meaning, Form, and Use in Context: Linguistic Applications*. Proceedings of GURT84. Washington: Georgetown University Press, 1984, pp. 11–42.

[Hul02] Joris Hulstijn. "Issues and awareness". In: *Sinn & Bedeutung VI, Proceedings of the Sixth Annual Meeting of the Gesellschaft für Semantik*. Ed. by Graham Katz, Sabine Reinhard, and Philip Reuter. University of Osnabrück. 2002.

[Kam78] Hans Kamp. "Semantics versus pragmatics". In: *Formal Semantics and Pragmatics*. Ed. by F. Guenthner and S. J. Schmidt. Springer, 1978.

[Kri07] Manfred Krifka. "Approximate interpretation of number words: A case for strategic communication". In: *Cognitive foundations of interpretation*. Ed. by Gerlof Bouma, Irene Kräer, and Joost Zwarts. Koninklijke Nederlandse Akademie van Wetenschapen, 2007, pp. 111–126.

[Lan84] Fred Landman. "Data semantics for attitude reports". In: *Varieties Of Formal Semantics: Proceedings of the fourth Amsterdam Colloquium, September 1982*. Ed. by Frank Veltman and Fred Landman. Groningen-Amsterdam Studies In Semantics. Reprinted as chapter 3 of [Lan86]; page numbers refer to this edition. 1984, pp. 193–218.

[Lan86] Fred Landman. "Towards a Theory of Information: The Status of Partial Objects in Semantics". PhD thesis. Universiteit van Amsterdam, 1986.

[Lew69] David K. Lewis. *Convention*. Cambridge: Harvard University Press, 1969.

[Lew73] David K. Lewis. *Counterfactuals*. Harvard University Press, 1973.

[Lew76] David K. Lewis. "Probabilities of conditionals and conditional probabilities". In: *The Philosophical Review* 85 (1976), pp. 297–315.

[Lew79] David K. Lewis. "Scorekeeping in a language game". In: *Journal of Philosophical Logic* 8 (1979), pp. 339–359.

[Lew81] David K. Lewis. "Ordering semantics and premise semantics for counterfactuals". In: *Journal of Philosophical Logic* 10.2 (May 1981), pp. 217–234.

[Lew96] David K. Lewis. "Elusive knowledge". In: *Australasian Journal of Philosophy* 74.4 (Dec. 1996), pp. 549–567.

[Low90] E. J. Lowe. "Conditionals, context, and transitivity". In: *Analysis* 50 (1990), pp. 80–87.

[MR94] Salvatore Modica and Aldo Rustichini. "Awareness and partitional information structures". In: *Theory and Decision* 37 (1994), pp. 107–124.

[MS82] Paul Milgrom and Nancy Stokey. "Information, trade and common knowledge". In: *Journal of Economic Theory* 26 (1982), pp. 17–27.

[McC80] J. McCarthy. "Circumscription — A form of non-monotonic reasoning". In: *Artificial Intelligence* 13 (1980), pp. 27–39.

[Mer99] Arthur Merin. "Information, relevance, and social decisionmaking: Some principles and results of decision-theoretic semantics". In: *Logic, Language, and Information*. Ed. by L. Moss, J. Ginzburg, and M. de Rijke. Vol. 2. CSLI Publications, 1999, pp. 179–221.

[Mos07] Sarah Moss. "On the pragmatics of counterfactuals". Unpublished manuscript, MIT. 2007.

[Mus95] Reinhard Muskens. *Meaning and partiality*. Studies in logic, language, and information. CSLI Publications, 1995.

[Par01] Prashant Parikh. *The Use of Language*. Stanford, California: CSLI Publications, 2001.

[Par92] Prashant Parikh. "A game-theoretic account of implicature". In: *Proceedings of* TARK IV. Ed. by Yoram Moses. 1992, pp. 85–93.

[RS04] Robert van Rooij and Katrin Schulz. "Exhaustive interpretation of complex sentences". In: *Journal of Logic, Language and Information* 13 (2004), pp. 491–519.

[RS61] H. Raiffa and R. Schlaifer. *Applied Statistical Decision Theory*. MIT Press, 1961.

[Roo03a] Robert van Rooy. "Asserting to resolve decision problems". In: *Journal of Pragmatics* 35 (2003), pp. 1161–1179. DOI: 10.1016/S0378-2166(02)00186-8.

[Roo03b] Robert van Rooy. "Questioning to resolve decision problems". In: *Linguistics and Philosophy* 26 (2003), pp. 727–763.

[Roo04] Robert van Rooy. "Signalling games select Horn strategies". In: *Linguistics and Philosophy* 27 (2004), pp. 493–527.

[Sch07] Katrin Schulz. "Minimal Models in Semantics and Pragmatics: Free choice, exhaustivity, and conditionals". PhD thesis. Universiteit van Amsterdam, 2007.

[Sch] Katrin Schulz. "If you wiggle A, the B will change". To appear in *Synthese*.

[Sta68] Robert Stalnaker. "A Theory of conditionals". In: *Studies in Logical Theory*. Ed. by Nicholas Rescher. Vol. 2. Oxford University Press, 1968, pp. 98–112.

[Sta84] Robert Stalnaker. *Inquiry*. MIT Press, 1984.

[Sta99] Robert Stalnaker. *Context and Content: Essays on intentionality in speech and thought*. Oxford cognitive science series. Oxford University Press, 1999.

[Tic76] Pavel Tichý. "A counterexample to the Stalnaker-Lewis analysis of counterfactuals". In: *Philosophical Studies* 29 (1976), pp. 271–273. DOI: 10.1007/BF00411887.

[Vel05] Frank Veltman. "Making counterfactual assumptions". In: *Journal of Semantics* 22 (2005), pp. 159–180. DOI: 10.1093/jos/ffh022.

[Vel81] Frank Veltman. "Data semantics". In: *Formal methods in the study of language (part 2)*. Ed. by Jeroen Groenendijk, Theo Janssen, and Martin Stokhof. Reprinted with minor alterations in [GJS84]; page references are to this edition. The Mathematical Centre, 1981, pp. 541–566.

[Vel85] Frank Veltman. "Logics for Conditionals". PhD thesis. Universiteit van Amsterdam, 1985.

[Vel86] Frank Veltman. "Data semantics and the pragmatics of indicative conditionals". In: *On Conditionals*. Ed. by Elizabeth Closs Traugott et al. Cambridge University Press, 1986, pp. 123–168.

[Vel96] Frank Veltman. "Defaults in update semantics". In: *Journal of Philosophical Logic* 25 (1996), pp. 221–261.

[War81] Ken Warmbrōd. "Counterfactuals and substitution of equivalent antecedents". In: *Journal of Philosophical Logic* 10.2 (May 1981), pp. 267–289.

[Wil08] J. Robert G. Williams. "Conversation and conditionals". In: *Philosophical Studies* 138.2 (Mar. 2008), pp. 211–223.

[Wit69] Ludwig Wittgenstein. *On Certainty*. Ed. by G. E. M. Anscombe and G. H. von Wright. Blackwell, 1969.

[Yal08] Seth Yalcin. "Modality and Inquiry". PhD thesis. MIT, 2008.

Abstract/Samenvatting

Abstract

This dissertation applies the notion of UNAWARENESS to problems of formal semantics and pragmatics. Unawareness is an epistemic attitude that has recently raised a lot of interest in epistemic logic circles, as well as in what we might call "formal epistemic economics". Informally it is closely related to INATTENTION: an (epistemic) agent may attend to possibilities (that is, consciously represent them and reason about them in deliberation) or be unaware of them (not give them conscious representation; not have them play any role in deliberation). While unawareness implies lack of knowledge, it differs from previous notions of uncertainty in its formal and conceptual properties; most importantly, an agent unaware of some proposition p does not know that p, but he also does not know *that* she does not know that p.

In Chapter 1 I describe the informal notions of unawareness and inattention and give some examples suggesting their applicability to formal semantics and pragmatics; these use a notion of ASSUMPTION that does not feature in the existing formal theories. I give a short survey of existing models, and argue that none such is appropriate for the linguistic problems; the rest of the dissertation tries to fill the resulting gap in the market.

In Chapter 2 I introduce the formal terminology of unawareness/inattention and assumption, and a simple logic with a static possible-worlds semantics. Chapter 3 gives a dynamic semantics, allowing us to describe changes in awareness, and argues that this is the most relevant framework for linguistic applications of the notions. Chapter 4 is a case study, applying unawareness to so-called "Sobel sequences", a long-standing puzzle concerning the semantics of counterfactuals.

Chapter 5 takes a different tack, developing a decision-theoretic apparatus enhanced with a representation of unawareness and assumption. The aim is to extend the range of decision-theoretic pragmatics, which describes various forms of pragmatic inference as rational behaviour according to decision-theoretic norms, to cover unawareness phenomena.

Chapter 6 gives a rather different unawareness model, based on data semantics. This captures various kinds of defeasible inference which owe their defeasible nature to unawareness (typically inference from evidence to "must"-statements, which are only justified under limited awareness of the domain of possibility).

Finally, Chapter 7 summarises the approach here presented and offers some speculation about possible future extensions of the ideas.

Samenvatting

Dit proefschrift past het begrip ONBEWUSTZIJN toe op problemen uit de formele semantiek en pragmatiek. Onbewustzijn ("unawareness") is een epistemische houding die recentelijk veel interesse gewekt heeft zowel in de epistemische logica als binnen wat wij de "formele epistemische economie" zouden kunnen noemen. Informeel gesproken is het begrip nauw verbonden met ONOPLETTENDHEID: een (epistemische) agent kan op bepaalde mogelijkheden letten (dat wil zeggen, ze bewust voorstellen en gebruiken bij het redeneren), of zich er niet van bewust zijn (oftewel ze niet bewust representeren; ze spelen dan geen rol bij het redeneren). Hoewel onbewustzijn de afwezigheid van kennis impliceert, verschilt het toch in zijn formele en conceptuele eigenschappen van eerdere noties van onzekerheid; essentieel is, dat een agent die zich niet bewust is van een propositie p, niet weet dat p, maar ook niet weet *dat* ze p niet weet.

In hoofdstuk 1 beschrijf ik de informele begrippen van onbewustzijn en onoplettendheid en geef ik enkele voorbeelden die suggereren hoe die toegepast kunnen worden op de formele semantiek en pragmatiek; daarbij speelt het begrip AANNAME een belangrijke rol, dat nog niet in bestaande theorieën voorkomt. Ik geef een kort overzicht van de bestaande modellen, en ik argumenteer dat die niet geschikt zijn voor de taalkundige problemen; de rest van het proefschrift probeert het ontstane 'gat in de markt' op te vullen.

In hoofdstuk 2 wordt de formele terminologie van onbewustzijn/onoplettendheid en aannames geïntroduceerd en ook een eenvoudige logica met een statische semantiek van mogelijke werelden ontwikkeld. Hoofdstuk 3 geeft een dynamische semantiek waarmee veranderingen in het bewustzijn beschreven kunnen worden, en stelt dat dit het meest relevante kader is voor taalkundige toepassingen van de begrippen. Hoofdstuk 4 beschrijft een casus, waarin onbewustzijn toegepast wordt op zogenaamde "Sobel sequences", een lang bestaande puzzel over de semantiek van counterfactuals.

Hoofdstuk 5 neemt een andere wending, met het ontwikkelen van een beslis-theoretische apparatuur uitgebreid met representaties van onbewustzijn en aannames. Het doel is het bereik van de beslis-theoretische pragmatiek, die verschillende vormen van pragmatische inferenties beschrijft als rationeel gedrag volgens beslis-theoretische normen, uit te breiden teneinde onbewustzijnsverschijnselen te vatten.

Hoofdstuk 6 geeft een ander model van onbewustzijn, gebaseerd op datasemantiek. Dit model beschrijft enkele soorten defeasible gevolgtrekkingen die hun defeasible karakter te danken hebben aan onbewustzijn (meestal gaat het om inferenties vanuit bewijs naar "moet"-beweringen; die zijn slechts te rechtvaardigen onder beperkte kennis van het domein van mogelijkheden).

Tot slot geeft hoofdstuk 7 een samenvatting van de benadering die hier gepresenteerd wordt en bied wat speculatie over mogelijke toekomstige uitbreidingen van de ideeën.

Titles in the ILLC Dissertation Series:

ILLC DS-2001-01: **Maria Aloni**
Quantification under Conceptual Covers

ILLC DS-2001-02: **Alexander van den Bosch**
Rationality in Discovery - a study of Logic, Cognition, Computation and Neuropharmacology

ILLC DS-2001-03: **Erik de Haas**
Logics For OO Information Systems: a Semantic Study of Object Orientation from a Categorial Substructural Perspective

ILLC DS-2001-04: **Rosalie Iemhoff**
Provability Logic and Admissible Rules

ILLC DS-2001-05: **Eva Hoogland**
Definability and Interpolation: Model-theoretic investigations

ILLC DS-2001-06: **Ronald de Wolf**
Quantum Computing and Communication Complexity

ILLC DS-2001-07: **Katsumi Sasaki**
Logics and Provability

ILLC DS-2001-08: **Allard Tamminga**
Belief Dynamics. (Epistemo)logical Investigations

ILLC DS-2001-09: **Gwen Kerdiles**
Saying It with Pictures: a Logical Landscape of Conceptual Graphs

ILLC DS-2001-10: **Marc Pauly**
Logic for Social Software

ILLC DS-2002-01: **Nikos Massios**
Decision-Theoretic Robotic Surveillance

ILLC DS-2002-02: **Marco Aiello**
Spatial Reasoning: Theory and Practice

ILLC DS-2002-03: **Yuri Engelhardt**
The Language of Graphics

ILLC DS-2002-04: **Willem Klaas van Dam**
On Quantum Computation Theory

ILLC DS-2002-05: **Rosella Gennari**
Mapping Inferences: Constraint Propagation and Diamond Satisfaction

ILLC DS-2002-06: **Ivar Vermeulen**
A Logical Approach to Competition in Industries

ILLC DS-2003-01: **Barteld Kooi**
Knowledge, chance, and change

ILLC DS-2003-02: **Elisabeth Catherine Brouwer**
Imagining Metaphors: Cognitive Representation in Interpretation and Understanding

ILLC DS-2003-03: **Juan Heguiabehere**
Building Logic Toolboxes

ILLC DS-2003-04: **Christof Monz**
From Document Retrieval to Question Answering

ILLC DS-2004-01: **Hein Philipp Röhrig**
Quantum Query Complexity and Distributed Computing

ILLC DS-2004-02: **Sebastian Brand**
Rule-based Constraint Propagation: Theory and Applications

ILLC DS-2004-03: **Boudewijn de Bruin**
Explaining Games. On the Logic of Game Theoretic Explanations

ILLC DS-2005-01: **Balder David ten Cate**
Model theory for extended modal languages

ILLC DS-2005-02: **Willem-Jan van Hoeve**
Operations Research Techniques in Constraint Programming

ILLC DS-2005-03: **Rosja Mastop**
What can you do? Imperative mood in Semantic Theory

ILLC DS-2005-04: **Anna Pilatova**
A User's Guide to Proper names: Their Pragmatics and Semanics

ILLC DS-2005-05: **Sieuwert van Otterloo**
A Strategic Analysis of Multi-agent Protocols

ILLC DS-2006-01: **Troy Lee**
Kolmogorov complexity and formula size lower bounds

ILLC DS-2006-02: **Nick Bezhanishvili**
Lattices of intermediate and cylindric modal logics

ILLC DS-2006-03: **Clemens Kupke**
Finitary coalgebraic logics

ILLC DS-2006-04: **Robert Špalek**
Quantum Algorithms, Lower Bounds, and Time-Space Tradeoffs

ILLC DS-2006-05: **Aline Honingh**
The Origin and Well-Formedness of Tonal Pitch Structures

ILLC DS-2006-06: **Merlijn Sevenster**
Branches of imperfect information: logic, games, and computation

ILLC DS-2006-07: **Marie Nilsenova**
Rises and Falls. Studies in the Semantics and Pragmatics of Intonation

ILLC DS-2006-08: **Darko Sarenac**
Products of Topological Modal Logics

ILLC DS-2007-01: **Rudi Cilibrasi**
Statistical Inference Through Data Compression

ILLC DS-2007-02: **Neta Spiro**
What contributes to the perception of musical phrases in western classical music?

ILLC DS-2007-03: **Darrin Hindsill**
It's a Process and an Event: Perspectives in Event Semantics

ILLC DS-2007-04: **Katrin Schulz**
Minimal Models in Semantics and Pragmatics: Free Choice, Exhaustivity, and Conditionals

ILLC DS-2007-05: **Yoav Seginer**
Learning Syntactic Structure

ILLC DS-2008-01: **Stephanie Wehner**
Cryptography in a Quantum World

ILLC DS-2008-02: **Fenrong Liu**
Changing for the Better: Preference Dynamics and Agent Diversity

ILLC DS-2008-03: **Olivier Roy**
Thinking before Acting: Intentions, Logic, Rational Choice

ILLC DS-2008-04: **Patrick Girard**
Modal Logic for Belief and Preference Change

ILLC DS-2008-05: **Erik Rietveld**
Unreflective Action: A Philosophical Contribution to Integrative Neuroscience

ILLC DS-2008-06: **Falk Unger**
Noise in Quantum and Classical Computation and Non-locality

ILLC DS-2008-07: **Steven de Rooij**
Minimum Description Length Model Selection: Problems and Extensions

ILLC DS-2008-08: **Fabrice Nauze**
Modality in Typological Perspective

ILLC DS-2008-09: **Floris Roelofsen**
Anaphora Resolved

ILLC DS-2008-10: **Marian Counihan**
Looking for logic in all the wrong places: an investigation of language, literacy and logic in reasoning

ILLC DS-2009-01: **Jakub Szymanik**
Quantifiers in TIME and SPACE. Computational Complexity of Generalized Quantifiers in Natural Language

ILLC DS-2009-02: **Hartmut Fitz**
Neural Syntax

ILLC DS-2009-03: **Brian Thomas Semmes**
A Game for the Borel Functions

ILLC DS-2009-04: **Sara L. Uckelman**
Modalities in Medieval Logic

ILLC DS-2009-05: **Andreas Witzel**
Knowledge and Games: Theory and Implementation

ILLC DS-2009-06: **Chantal Bax**
Subjectivity after Wittgenstein. Wittgenstein's embodied and embedded subject and the debate about the death of man.

ILLC DS-2009-07: **Kata Balogh**
Theme with Variations. A Context-based Analysis of Focus

ILLC DS-2009-08: **Tomohiro Hoshi**
Epistemic Dynamics and Protocol Information

ILLC DS-2009-09: **Olivia Ladinig**
Temporal expectations and their violations

ILLC DS-2009-10: **Tikitu de Jager**
"Now that you mention it, I wonder…": Awareness, Attention, Assumption

ILLC DS-2009-11: **Michael Franke**
Signal to Act: Game Theory in Pragmatics

ILLC DS-2009-12: **Joel Uckelman**
More Than the Sum of Its Parts: Compact Preference Representation Over Combinatorial Domains