A Demand of Reason:

Dependence in Logic and Probability

MSc Thesis *(Afstudeerscriptie)*

written by

Nathaniel Forde
(born April 6th, 1987 in Dublin)

under the supervision of **Dr. Sonja Smets** and **Prof.dr Martin Stokhof**, and submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam.*

Date of the public defense:	**Members of the Thesis Committee:**
September 2nd, 2013	Dr Maria Aloni
	Dr. Ing. Robert van Rooij
	Dr Sonja Smets
	Prof.dr Martin Stokhof
	Prof.dr Frank Veltman

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

Abstract

In this thesis we shall argue on two levels (i) we shall elaborate and criticise the traditional models of formal epistemology (i.e. the standard Bayesian and Hintikka-style possible world models) and (ii) we shall show that there is a theory of explanation which incorporates the influence of both formal epistemology and formal ontology. In particular we shall develop an account of explanation based on the explicit logics of dependency relations (e.g. Kit Fine's logic of grounding and Pearl and Halpern's logic of causal dependence) and the differing species of justification logic (*a la* Sergei Artemov) which are defined in terms of each such dependency relation. The incorporation of this plurality of justification logics in formal epistemology allows us to deal with some of the problems afflicting the traditional models of formal epistemology, and construct a theory of explanation which directly encodes the relevance of the *explanans* to the *explanandum.* Finally, this theory of explanation is opposed to van Fraassen's account explanation and defended as both superior and complementary, thereby motivating further work at the intersection of formal epistemology and formal ontology.

Acknowledgements

"I get by with a little help from my friends". In particular, this thesis would never have begun were it not for Paal - who taught by example. In 2011 I resolved to "try with a little help from my friends" this thing called logic. Looking back "when I was younger, so much younger than today" I can't say that "I never needed anybody's help in any way." The first semester in Amsterdam was tough. Learning logic often seemed like trying to "speak double dutch to a real double duchess." To my friends, I have to say "I do appreciate you being round" you helped me "get my feet on the ground." This can't be overstated, for the "greatest thing you can ever do/ Is trade a smile with someone who's blue." Even now, after we've "come together" I feel oddly grateful for the "hours that the clock cannot define" spent in the seemingly shared agony of the MoL. But beyond the MoL, the support of my family and friends back home made this adventure possible - "I'd be lost if not for you/ And you know it's true." In particular I owe my brother Jamie; his ambition is always an inspiration. There is not enough room to thank my parents in full. Their tolerance is appreciated, especially given that "[on] this obstacle course/... it [often] seems I'm getting nowhere" trying to "ride the concept of the horse" while ignoring practical alternatives. Finally, and most directly, this thesis would not have been finished without Sonja and Martin, who kindly agreed to jointly supervise my fumbling efforts. That we managed to distil anything at all from my initial enthusiastic ramblings is a testament to their patience and prodding. Their insightful questions more often than not broke the"paralysis of analysis" which prevents pen from contacting paper. I hope you both find the final product somewhat satisfying and not entirely false - its completion almost made me "insane in da brain."

Thanks goes, in no particular order, to: Jmo, Hans (for the Amstel), Fenner, Aadil, Tanmay (for the route home), Hugo, Ásgeir, Ryan, Alwin, Paula (especially for Induction), Marie, Gianluca, Giovanni, Iris, Helena, Johannes, Alex, Pedro, Roos, Cian, Alaina, Anthony, Maja, Callum, George (Judy), Katya, Dieuwke, Femke, Rob, Roosmarijn (especially for set theory), Nikhil, Aafke, Andreea, Erik, Vlasta, Max (especially for modal logic), Zoe, Masa, Sanne, Adam, Will, Sebastian, Justin, Gabor, Aybuke (for reminding me), Azarah, Jennifer, Kletia, Zhiguang, Bora, Tanvi, Alan, Jim (Katherina), Julia, Finn, Sean, Eamonn, Ciaran, Seb, Tanja (for everything!) and too many others (inside and outside the ILLC) to mention. Despite all this support, all below errors are mine.

Contents

Preface

Reason makes this demand in accordance with the principle that if the conditioned is given, the entire sum of the conditions, and consequently the absolutely unconditioned... is also given. - Kant in *The Critique*[1]

Introduction

This thesis is best viewed as a polemic. The target of the polemic is certain kind of specialisation in philosophy; namely, formal epistemology. We seek to argue that current models in formal epistemology are plagued with problems. On a more positive note we suggest that the solution to these problems becomes clear if we simply stop treating the epistemic and doxastic notions as unanalysed primitives. We seek to unpack the notions of belief and knowledge in such a way so as to reveal the evidence which underlies our beliefs and knowledge. To do so we explicitly link formal epistemology with formal ontology. The broad moral of this thesis is simple:

> No formal epistemology without formal ontology.

Less enigmatically, we seek to elaborate the conditions under which reasonable belief arises. To do so we find that we have to draft principles of inference which respect the natural dependency structures between the occurrence of the events about which we reason. Ultimately we wish to show that the idealisations which underlie traditional models of knowledge and belief are inappropriate to cash out any notion of rational belief or knowledge. But by including these dependency structures we can suitably amend our models of belief to capture various species of rational inference. In particular this approach allows for the development of a theory of explanation in a dynamic epistemic setting. To develop this conclusion we argue in the following pattern.

Abstract Thesis Structure

Structurally the thesis is split into two sections. The first three chapters are spent initially on elaborating a problem for epistemology generally; namely underdetermination of rational belief, and secondly we develop two paradigmatic models of formal epistemology. The hope is that by establishing these models we can address the underdetermination problem. The idea here is that (a) one necessary condition for any model of formal epistemology is that the model be able to articulate the traditional problems of epistemology and that (b) a condition of adequacy for any model of formal epistemology is that it be able to address the traditional problems of epistemology. In other words, we ought be able to formally model the resolution of those problems. We first show that the traditional models of formal epistemology do not satisfy either constraint, hence we recommend an alteration of such models. On a positive note we draw an instructive moral from the failings of our epistemic doxastic models. With this moral in hand we turn to the second section (i.e the last three chapters) of our thesis.

[1] See either A409 or B436.

In this part of the thesis we aim to augment our epistemic doxastic models by the inclusion of structural dependency information. For instance, we shall focus in chapter five on how our knowledge of a certain kind of explanatory relation subsists on the observations of causal dependence. With this augmented model we show how we may resolve or dissolve the earlier underdetermination problems. Of the last three chapters, our fourth chapter is spent presenting an argument for why the inclusion of structural dependency information can play a positive role in our models of knowledge and belief. Our fifth chapter compounds this argument by giving an example of an epistemic doxastic model which is positively improved by the inclusion of structural dependency information. In our sixth and final chapter we set ourselves two tasks: on the one hand we aim to suggest the shape of future work in which we can develop a model of explanation as an update operation of suitably detailed information states. These suggestions are just that. On the other hand, we defend our postulation of real (extant) dependency structures against van Fraassen style objections to our brand of structural realism. We take our defence to be more than adequate. On this we conclude.

Detailed Thesis Structure

Chapter 1: Arguably True In this chapter we present Goodman's grue paradox as an underdetermination problem for the notion of rational belief. That is to say we use Goodman's argument to suggest that belief is always question begging and knowledge unattainable. The majority of this chapter is spent informally developing the notion of theory preference and choice, in such a way that there is reason to reject the grue-hypothesis. We draw on the work of Kevin Kelly and Peter Lipton in our discussion of theory preference and argue that underdetermination results pose a problem in so far as they motivate a search for non-obvious choice criteria. We conclude this chapter with a general discussion of underdetermination problems and the structure of their solution, drawing on the work of Clark Glymour. These observations play a motivating role in all which follows since the underdetermination problem provides a test of adequacy for our models of formal epistemology.

Chapter 2: Probably True In this chapter we seek to elaborate the grue problem in the standard Bayesian model of belief. The hope is that such an elaboration will be able to accommodate the kind of solution we glossed in the informal discussion of undetermination problems in the previous chapter. This fails. We then develop some criticisms of the Bayesian model, not least of which is the failure to cope adequately with underdetermination problems. We also consider the role of Bayesian theory as against some famous results in the psychology of reasoning drawing here on the work of L.A. Paul and Kahneman and Tversky. We conclude the chapter with a recognition of the limits of the Bayesian model, in particular the failure to capture the notion of when a statistical correlation encodes evidentiary relevance. The main point of this chapter is that the notion of belief is not well understood on a Bayesian setting. We follow van Fraassen's voluntarist position in this conclusion.

Chapter 3: Plausibly True In this chapter we shift to a qualitative model of belief and knowledge, motivated by the failings of the Bayesian model. We elaborate a Hintikka style analysis of knowledge and belief before pointing to some difficulties of these models. Again we try to show that these difficulties emerge from the assumption of a primitive unanalysed notions of knowledge and belief. We draw on the work of Sellars and Fitch to emphasise this point. These observations motivate our turn to Sergei Artemov's justification logics. We show the connection between justification logic and doxastic logic, before utilising this new setting to articulate the structure, and (a too formal) solution of the grue problem. Crucially, the structure of this solution suggests the need for analysing the notion of evidentiary relevance. To properly account for this moral we must examine the type of dependence relation which obtains between our premises and conclusions.

Chapter 4: True In this chapter we seek to show (a) how underdetermination problems underlie many of the issues which afflict both Bayesian and Hintikka style models of knowledge and belief, and (b) how the incorporation of structural dependencies information can be used to avoid such underdetermination problems, and (c) how such structural information can be used to define multiple notions of evidentiary relevance. We conclude this chapter by showing how to link Kit Fine's logic of grounding relations with a species of justification logic. In this way we hope to circumscribe a species of justification based on a particular kind of dependency relation. We deploy these considerations in the evaluation of Benacceraf's famous dilemma.

Chapter 5: Because In this chapter we elaborate a model of causal dependence discussed by Judea Pearl, and formalise the logic causal dependence defined by Pearl and Halpern. The idea is to show that the abstract discussion about the role of dependency relations in last chapter, can be made concrete. We apply this to the consideration of the grue-paradox, and show that the latter is resolved. We further argue with Pearl that the failure to incorporate such dependency relations in our models effectively cripples our ability to provide reasonable models of the world, and hence cripples our ability to draw reasonable inferences. This is not quite a reductio of the alternatives, but it does demonstrate the kind of impoverishment afflicting the standard models of formal epistemology.

Chapter 6: Arguably, arguably true In this chapter we sketch some ideas for the further development of a model of explanation as a species of information update, based upon the foregoing discussion of evidentiary relevance and justification logics. Perhaps more significantly we defend the species of structural realism we have hitherto elaborated. The nature of the defence is largely offensive. We attack van Fraassen's theory of constructive empiricism as an adequate of model scientific reasoning, thereby (we hope) motivating our modest postulation of effectively unobservable dependency structures.

Conclusion We conclude by situating our discussion amongst some existing models of explanation. Ultimately we recommend that our approach is both complementary of most such models, and an improvement on all. We credit this result to the explicit inclusion structural dependencies within our epistemic considerations. So we recommend further research in the intersection of formal epistemology and formal ontology.

How to (Quickly) Read this Thesis

The thesis is slightly longer than I had hoped due to the nature of the target. The characterisation of our target in chapters two and three was a necessary if perhaps tedious pre-requisite for the development of our argument. If you already agree that the standard Bayesian and Hintikka style models of knowledge and belief are somewhat too idealised, then you might be able to profitably skip from the end of chapter one to section 3.3. The crucial move here is that we criticise the traditional models of formal epistemology for treating knowledge and belief as unanalysed primitives. We recommend justification logic as a solution to this problem which allows us to represent conditions on which the emergence of belief and knowledge can be seen to be dependent. We then continue to argue that by focusing on different species of dependence we can lend plausibility to models of reasoning which do not seem to respect either the dictates of logic (i.e entailment) or probability (i.e correlation) traditionally construed. It is this observation which prompts our recommendation that formal epistemology is well complemented by a focus on formal ontology. We reason about a number of forms of dependence, and we do not reason well unless we are clear about the nature of these dependence relations. To defend this conclusion we argue that our attention to a number of distinct dependency relations can be seen to motivate a reasonable (i.e. defensible) theory of explanation. We then defend this theory against van Fraassen style worries.

If this all sounds eminently reasonable, non-trivial and beyond refutation feel free to stop reading now, otherwise please continue.

How to Read this Thesis

One page at a time. Thanks!

Chapter 1

Arguably True: Underdetermination Problems

I choose my friends for their good looks, my acquaintances for their good characters, and my enemies for their intellects. A man cannot be too careful in the choice of his enemies. - Oscar Wilde

1.1 Introduction: Goodman's Riddle

We shall argue that a particular kind of underdetermination problem admits a solution. Crucially, we shall treat underdetermination as a problem for rational choice, rather than a mere fact of theory construction. In this chapter we present the issue and in later chapters we treat it as a paradigmatic instance of a general form of problem we wish to solve.

The Problem

Assume that a law of the form $\forall$x(Ex $\rightarrow$ Gx) whether scientific or otherwise, is confirmed by its instances. So naturally, any well established record of a universal claim which accords with the facts, might be taken to confirm our hypothesis to a high degree. In particular, we might think that if every emerald hitherto observed has been seen to be green, the claim that all things, if they are emeralds are green, is true. If this law holds, then it supports counterfactual reasoning about emeralds. For then we know that for some particular jewel, if it were an emerald, it would be green. We might even be tempted to say that this law is an explanation for both our current information and our expectations. So clearly many issues depend on our ability to endorse such a law. However, there is a problem. Nelson Goodman argues[1] that we cannot exclusively endorse one law on the basis of our evidence. There are many mutually exclusive laws which are compatible with our evidence, and so the appeal to explanatory laws is fundamentally underdetermined. Why this law, and not another?

Illustration

By way of example observe that we may define a law to accord with our evidence but that differs from the intuitively agreeable hypothesis. Suppose that you encounter an advocate for the grue-hypothesis; that is the claim that $\forall$x(x is an Emerald, $\rightarrow$ x is Grue), where Grue is a colour that changes upon time-indexed instances of observation. In particular, our antagonist claims that all emeralds are grue just when all emeralds observed before the year 2015 are seen to be green, and all emeralds observed after the year 2015 are seen to be blue. As such our current information equally confirms the hypothesis that all emeralds are green, as much as it confirms our antagonist's claims. Fundamentally this becomes a problem of determining which theory provides the more plausible

[1] The classical statement of this argument is found in [42]

explanation. If you accept that plausibility rankings are arbitrary, then we have no non-arbitrary reason to prefer one hypothesis over the other. This comes to the view that all explanatory laws are equally viable. If you claim our ranking is non-arbitrary and provide a criteria by which to rank explanatory laws, then the sceptic can easily generate alternative criteria which results in the inverse ranking of explanatory hypotheses. So it appears that we have no motivation to accept any explanation or a theory of explanation. This is an intolerable sceptical thesis which seems to motivate the view that all explanatory laws are equally plausible so long as they are consistent with what we know.

Connected Problem

This problem repeats in a more obviously pernicious fashion if we consider counterfactual reasoning. Any counterfactual claim which states that ψ would be true, if it were the case that ϕ can only be judged true when we are in a position to know, putting it bluntly, that ϕ was relevant to the evaluation of ψ.[2] If we can't determine relevance, then we can't evaluate counterfactuals. Following Hume, this is also a clear statement of a problem which plagues any account of inductive reasoning. For if it's not clear over which counterfactual situations in which ϕ is relevant to ψ, it's not clear in which future states ϕ is relevant to the truth of ψ. In effect Goodman's argument prompts the conclusion that we impose our arbitrary individual standards of relevance on inductive and counterfactual reasoning. We do so because we wish to endorse the emerald-hypothesis and reject the grue-claim.

Discussion

It might be natural to suggest that there is always some law available to introspection which assures the connection between the antecedent ϕ and the consequent ψ. However, the formulation of this law requires that we specify (a) the compatibility or cotenability of our antecedent with our consequent, and (b) the necessary or plausible connection between the truth of the antecedent, in some specific situation, and the truth of the consequent. Given that the specification of (b) is arguably difficult to defend *a posteriori* for Humean reasons, we might hope it admits an *a priori* justification... unfortunately such a presumption is unwarranted in general. Worse, we are faced with the problem of circularity since the specification of (a) involves the elaboration of a counterfactual which ensures the contenability of the antecedent and consequent in all counterfactual situations. In other words, we need to say that if it were the case that ϕ, then ψ is never an unreasonable hypothesis. Usually we would spell out why ψ is a reasonable hypothesis with reference to some law or other which is assumed to hold over said counterfactual situation. Loops beckon.

For example the claim "If it were the case that your jewels were emeralds, then your jewels would be green" relies on the implicit premise that "if anything is an emerald, it would always be a green jewel". As such, no counterfactual can be understood without the understanding of counterfactuals; specifically the laws involved in our projections across counterfactual situations. This is either a circle or a regress which proves problematic if we seek to justify the laws we use. But do we even have a good grasp of the laws governing our projections? In other words, can we give an *a priori* specification of when a particular inductive generalisation is warranted? Can we inductively learn when our best information about emeralds is suitably secure so that it motivates a unrestricted projection of our candidate law across all possible situations? Goodman's riddle is an argument to the effect that we do not have a clear grasp of the validity of our projective estimates because we cannot say when a particular law is relevant to the assessment of a given hypothesis. This does not entail that we cannot argue for our preferences.

[2] We're not assuming any particular notion of relevance, but this could be fleshed out either in terms of probabilistic, causal relevance, or more liberally we could let relevance be an intransparent primitive.

Preliminary Conclusion

Two schematic solutions suggest themselves. Either adopt a primitive ranking of explanations and concede that our ranking of such explanations is not epistemologically transparent i.e. that we are ignorant of why the grue-hypothesis is implausible, or we can accept some such explanations as self-explanatory. The latter move is problematic because it leads to the mentioned regress, if the self-explanatory law is contestable. We can have explanatory chains which admit a benign regress, and take our preference ranking on explanations to be a non-transparent primitive. So for instance, we might explain a counterfactual claim in terms of the laws which underlie it, and the validity of laws in terms of the intuitively correct counterfactuals they support, while at the same time confessing ignorance about what makes a counterfactual intuitively correct. The process from which we reason on our ranking of explanatory laws is called the inference to the best explanation, in which we accept some explanatory laws just when we have determined it to be best of the available options. The open question is whether our best option is a true option? The assumption that we get it right, is often called the process of over-fitting, in so far as it risks a presumption that the world might not fit the laws we choose to project. Nevertheless, we reason by counterfactual considerations, and they seem to require projective laws. Furthermore over-fitting clearly works evidenced by the survival (to date) of our species; the riddle of induction is that we can't clearly say why it works.

You might think that over-fitting works because the laws which we choose are simply true. Not a radical idea admittedly, but at least now we have a tractable question. What reason do we have to believe that our law-like projections, and the theories in which they fit, are true? The hope is that if we can answer such a question we might be able to provide a motivation for the plausibility ranking of our possible explanations.

1.2 Truth Approximation

One reason to think our procedures of inference get it right is that the more effective our theory is, the more likely that it approximates the truth. Theory improvement occurs, so theory improvement makes it reasonable to think that our theories increasingly approximate truth.[3] Our theories improve because we incessantly make minor choices and adaptations to our theories in the aim of making them more explanatory or predictively accurate. We choose, for instance, to admit green emeralds and deny the grue-hypothesis. But in what sense is the grue-hypothesis worse than the alternatives? In this chapter we will argue that there are good reasons to reject the grue-hypothesis, because our theories increasingly approximate the truth and simplicity in theory choice can be used to rule out the grue-hypothesis.

1.2.1 Theory Choice and IBE

The process of inference to the best explanation is not a species of deduction. So by Goodman's reasoning any hitherto reliable appeal to a particular explanation will fall victim to one or other underdetermination scenario. This by now is familiar. Reliance on one or other explanation rests on an implicit premise regarding the uniformity of nature, but this uniformity cannot be projected into the future without begging the relevant question. Nevertheless, such projective estimates are regularly employed. This is in line with Goodman's observation that the problem of induction does not indicate defective reasoning so much as it highlights the mystery of what makes inductive reasoning valid. Why are some explanations projectable, and others not? Consider the following process. Upon observing a particular event $[\psi]$ for which they seek an explanation each agent

[3] For instance see the work of Niiniluoto in [68]

runs a comparative ranking of a short-list of theories determined to be viable by their background theory or beliefs. Now allow that this comparative ranking is effective or highly reliable.

Objection

Since the short-list of viable theories necessitates an exclusion of other theories we have no reason to think that the true theory will be in our short list. Taking this together, we might think that inference to the best explanation provides no assurance of truth whatsoever, because whatever short-list an agent generates cannot be thought of as privileged.

Response

Lipton[4] rejects this thought, on the basis that for any comparative ranking we need only insist that the short-list contain contrary explanatory theories. Since our process of comparative ranking is effective we now know that one of the contraries is more likely true. Hence, for any short-list we zero-in on the truth in a piece meal manner.

Objection

The short-list ranking is always apt to generate the simplest explanation, and the simplest explanation is always the easiest, so not necessarily the truth. Hence new theories are unlikely to be any more accurate than our background theory.

Response

This is clearly false. We are not trapped by the acceptance of one false belief to a steady increase in the falsity of our beliefs just because we seek conformity with our initial false belief. We don't merely seek conformity. So apart from being descriptively false, there is another problem with the view that comparative choice always favours the simplest explanations.[5]

Reflection

If you wished to maintain that scientists naively seek only simplicity or conformity, then you should reject the view that they can reliably rank any generated short-list of theories. Reliability is surely a guide to the truth, and simplicity alone is not. The reliability of the ranking mechanism entails the reliability of theory choice. Whereas the insistence on unnuanced simplicity (or some other ideological preference) contradicts the notion that our theory choice is truth-apt because it allows for the acceptance of a theory which is arbitrarily far from the truth.

Evidently we arrive at our background theory on the basis of a prior ranking, but if (as we've supposed) our method of comparative ranking is reliable then the background theory is supported as the result of a reliable mechanism. So we should expect to zero-in on the truth, because each stage of comparative analysis ensures increased accuracy by the iterative winnowing of poor choices. This motivates the following claim:

> If a scientist knows her method of ranking is reliable, then she is also in a position to know that her background is probably true, which entails she is capable of absolute evaluation....So the initially plausible idea that scientists might be completely reliable rankers yet arbitrarily far from the truth is an illusion.[6]

[4] [64]

[5] Even accepting that there is a phenomenon of confirmation bias, this by no means entails that we seek to confirm our beliefs by the simplest methods. Consider the exponential growth of complexity accompanying questionable conspiracy theories.

[6] [64]pg158.

Returning to the grue problem, we can see that this argument suggests that projectable explanations are required to be in conformity with some details of our background theory. This is just question begging if we mean that the projectable explanatory laws are simply those which are a generalisation of our hitherto observed evidence. In the grue-case, such a move is clearly too simple. However, we can seek to exclude the grue-case, not on the basis of it's unintuitive nature, but rather as a result of a broader categorical restriction. So you might suggest that our theory construction procedures have winnowed the notion of time indexed colour change from plausibility. This does not require that conformity with our accepted beliefs is the only or over-riding factor in theory choice. But where experience attests to the static nature of colour-properties, the burden of proof surely rests with the grue-advocate, especially if grue is the unique time-indexed colour. So at least we have no reason to accept the grue-hypothesis. This constitutes some progress.

1.2.2 Simplicity: Reconsidered

If we consider more general objections to the view that our theories approximate truth, they seem to require that we reject the assumption that we are any good at comparative ranking of theories, but this collapses into radical scepticism which is at least false if not also incoherent. A particular objection to this view questions the effect of simplicity as a criterion for theory choice. Above we saw a claim that the over-riding concern for simplicity will lead us astray, in such a way that we cannot expect to converge on the truth. But is this true?

There is undoubtedly a role for simplicity to play in the adjudication of preference, for as Kevin Kelly observes

> Ockham's razor is not a bloodless, formal rule that must be learned - it has a native visceral grasp on credence...Empirical simplicity is more than mere notational brevity - it implies such red blooded considerations as explanatory power, unity, independent confirmable principles and severe testability.[7]

That said, Kelly argues convincingly that simplicity as the sole criteria for preference involves a viscous circularity of reasoning. Ockham's razor cuts cleanly, but thankfully it is not the only instrument available to us. This is just as well, for as Kelly argues we cannot suppose that simplicity is a short term guide to truth without presupposing the inherent simplicity of nature. The claim that privileging simplicity in all choices will ensure that we shall eventually converge on the truth after incessantly running through all options repeatedly until our theory comes to fit the data, is a very weak claim. Any choice function based on nearly any criterion will eventually converge to truth in the ideal limit of inquiry. But in each of these cases we would not think that the preference-rankings prompted by these singular choice criteria are reliable. For effective choice we need to augment our decision criteria.

Effective Test on Theory choice

We might hope that a choice function converges on the truth just so long as there is an accurate test procedure for after choices are made, such that failure of the test prompts renewed choice and explicit rejection of the failed theory. In effect Kelly argues[8] for exactly this, when he states that the use of Ockham's razor in empirical circumstances leads efficiently to the truth because the range of any empirical inquiry is finite due to our restricted focus over the class of possible theories. The idea is that our choices must eventually accord with the truth fully because a scientist needs to ensure that any new information is consistent with our old (presumed true) information.[9] At each stage of inquiry there is check for the simplest theory compatible with the accrued experience

[7] Section 1 of "Simplicity, Truth and Probability" in [11] forthcoming.

[8] Sections 6 and 7 of the aforementioned essay in [11] forthcoming

[9] By the *stalwartness* condition of rational thought.You continue to defend what you know to be true.

of an agent. Accrued experience brings with it new observations which are taken to be true, and since consistency is an effective method of test, we can see that if the world is logically consistent, and our prior information was correct then Ockham's razor in empirical situations converges to the truth.

Of course, if our prior information was achieved by an application of Ockham's razor, we are back to square one. But since our methods of decision are based on more than considerations of brute simplicity, we need not worry overly much. Reasoning from true premises with a concern for simplicity is a reliable strategy so long as the truth of our premises wasn't established by appeal to Ockham's razor.[10] Hence Lipton's argument is correct, concern for simplicity alone allows that we can develop a theory which is arbitrarily far from the truth. However, for any effective method of decision, we can expect to converge on the truth after successive stages of inquiry.

The Grolour Argument

Now suppose that before the grue advocate there was a Grolour advocate. The Grolour hypothesis is the statement that all colours are time indexed. That is, every conceivable shade has a time parameter - after which the shade shifts by an observable degree anywhere the colour was instantiated. More precisely, allow that the infinite number shades were arrayed so that from the beginning of time objects have been shifting their shades every hour. Now the Grolour hypothesis could be refuted if we were to array objects of every shade and note the total failure to change colour. No one has had the time to bother or the relevant paints. Nevertheless no one is a Grolour advocate any longer. Why?

Take the needless complexity of the theory to be an indicator of it's falsity, and cite Ockham's razor as our criterion of choice, then we have a reason to reject the Grolour hypothesis. In particular, note that the Grolour hypothesis makes all our intuitive colour terms derivative constructs from the primitive vocabulary of time-indexed colours. If this is not reason enough, observe that if the grue-hypothesis is reasonable, the Grolour hypothesis would prompt innumerable specific grue-like hypotheses. In these situations you can imagine some enterprising young individual who predicts a *specific* colour change for the hour in question. The likelihood of predictive success is infinitesimal (even if the Groulor hypothesis were true!), and so any attempt to support the Grolour hypothesis by grue-like claims will (I'll bet) fail. However, the reverse is true if we argue against the Grolour hypothesis. Each prediction of a static colour instantiation has been borne out by the testimony of experience.[11] Now note that the grue-advocate is arguably committed to the Grolour hypothesis. By what standard could they deny it? If they can offer a reason for restricting change-parameters to blue/green, we can surely cite the same reason for denying even that! If they accept the Grolour hypothesis they have a poorly confirmed theory, even if the specific grue-hypothesis is equally well confirmed as the alternative green-hypothesis. As such we should reject the grue theory because either it entails a falsity or is itself arbitrary and question begging.

[10] In Kelly's setting, this reasoning is depicted as an inquiry problem (K, Q) which asks us to use the information we know K about observable effects in the world to choose an answer to a particular question Q which is a partition of optional theories and the empty set depicted as "?". Ockham's razor is a method M which says always choose the simplest theory in Q. But if we set the observable *stream of experience* to be empty, or deny that the true set of effects is in the observed set of effects K, then there is little reason to think Ockham's method converges on the truth. Either M would be a function M:K $\mapsto$?, modelling a state of constant indecision or M:K $\mapsto$ T_i for some arbitrary incorrect but simple theory.

[11] Realistically in both cases we have a finite resource bound on testing the claims, this resource can be construed (in orders of cynicism) in terms of time, patience, or the minimal number of experiments required for publication. In any case the grue-like claims will fare worse that the static-colour hypotheses.

1.2.3 Simple Logic: Toward Formalising Simplicity

We have elaborated some intuitive reasons to reject the grue-hypothesis but the problem admits treatment in a more formal setting if Ockham's razor applies. For instance if we could conclude with Kelly that:

> There is a structural sense in which constantly coloured worlds are simpler than worlds of changing colour...[I]n constantly coloured worlds, Nature eternally "reserves the right" to present an "anomaly" later by exhibiting a colour change, but in changing colour worlds she eventually exercises her right to change the colour and never gets an opportunity to make the world appear constantly coloured thereafter.[12]

However, to make this sense of structural simplicity explicit we need to delve further into a slightly more formal setting. In this section we examine a discussion of simplicity related in an widely circulated but unpublished manuscript [52] as the discussion is explicitly linked to the grue-paradox. [13]

The Setting: An **empirical problem** can be thought of as a question Q and a set of presuppositions K. The question can be conceived of as a partition of mutually incompatible theories $\{T_1, ...T_n\}$ i.e. potential answers. Our presuppositions are assumed to be consistent, and each theory determines a unique **correct answer** to every question. Given the empirical nature of the problem we can assume access to an constant **stream of experience**. This experience should be thought of as a series of discrete stages of informational update; construed either as the passing of time or successive stages of inquiry. A **method** is a function which maps each finite stream of experience compatible with the problem's presuppositions to a potential answer. In this setting a potential answer is either one amongst the partition of available theories or the state of ignorance "?". A method **solves** our problem just when it ensures that answer eventually stabilises or **converges** to the correct answer.

Every **method** has a resource bound (which can be finite or infinite), we say that a problem is solved efficiently or has an **efficient solution** just when the problem (and any sub-problems) are solved with the least achievable resource expenditure. The resource expenditure is here considered in terms of poor choices, and required retractions over the length of inquiry. So if for a method we need to move through n choices based on an increasingly large set of presuppositions to reach the convergence point (i.e the true theory), with k retractions (denoted k$_r$) than our method is n-k$_r$ **retraction efficient**. So our problem is efficiently solvable if our resource bound is greater than or equal to n-k$_r$. Of course the resource bound is the number of further steps available to us at any stage of inquiry.

The Task: Now recall that we wish to use Ockham's razor as our choice criterion to distinguish between theories. Let **Ockham's razor** be defined in terms of a problem relative to an ordinal ranking of simplicity. Which is to say, that the simplicity of a theory is to be defined (in terms of potentially anomalous parameters or free variables) against the background presuppositions involved in the statement of our empirical problem. Hence Ockham's razor is a choice method used to ensure that we choose the simplest theory given our information. We state the main result of Kelly's paper and then expand on some of these notions.

[12] [52]pg3

[13] The manuscript was downloaded from www.philpapers.org, but it can also be found here: www.andrew.cmu.edu/user/kk3n/homepage/kelly. A more systematic view of Kelly's work can be found in [53]

Proposition: (Ockham) *If a method solves a problem retraction efficiently then the method complies with Ockham's razor in the problem.*[14]

With the notion of simplicity so defined you might wonder how we could apply it?

An Application

Recall our grue-hypothesis. If the goal is to maintain simplicity and avoid retraction errors, then we are justified in maintaining that the world is constantly coloured. For if we are wrong, and Nature admits an "anomaly" we need only retract our error once. However, if we concede to the grue-advocate and the hypothesis fails we risk having committed two retraction errors. Of course we need our situation to be empirical to observe errors which would prompt retraction, but as the grue-case exemplifies we can advocate against theories which do violate the possibility of observational confirmation on the basis that they ignore Ockham's razor. For suppose that the grue-hypothesis is amended to predict a colour shift after the heat death of our sun, presuming we have exited the stage some time before, we could not by observation confirm or deny this grue-hypothesis. However, we can argue against its plausibility simply because it exposes us to a greater error potential than if we rest content and think of the world as constantly coloured.

Since every method is construed as a question answering mechanism we say that **Ockham's Answer** is determined by a function s: Q $\mapsto$ On, which assigns to each member of Q an ordinal ranking, so that the simplest answer is the theory which has the least degree of complexity i.e. has lesser potential anomalies than any other theory and is otherwise compatible with our presuppositions. Of course since an anomaly is defined as a deviation from any uniformity recorded in our stream of experience, this means that theories with fewer anomalies respective to our presuppositions are deemed simpler and therefore preferable. The number of anomalies which need to be confirmed or denied determine the length of our inquiry and the resources available determine the maximal length of our inquiry. In other words, Ockham's answer is the theory T_i such that the resource cost of checking T_i is less than the resources available. The process of checking can be thought of as the expenditure of time waiting for the potentially anomalous parameters of our theory to resolve themselves. In other words the simplicity ranking of our theory, given our presupposition is less than that of all competing theories, because it contains less potentially anomalous parameters. **Ockham's Razor** is the method that takes Ockham's answer to always be the best choice under the resource bound.

Note that on this measure of complexity we treat theories as being possessed of objectively measurable degrees of potentially anomalous parameters. The assumption is plausible only if Q partitions the state space in such a way that each theory is closed under logical entailment so that all potentially anomalous parameters are the only variable elements of each theory. This seems tolerably reasonable. However it is much less clear how, once those variable parameters have been allocated, we evaluate the probability of whether these parameters will be uniform with our acquired evidence and stream of experience or not. Unfortunately, this is exactly the grue problem, we are not *a priori* in a position to judge the grue-hypothesis less probable than the natural alternative. Seen this way Kelly's argument boils down to making a methodological choice for preferring theories with lesser retraction potential. However such a claim would not deter the sceptic, because the sceptic claims that your ability to detect error doesn't work either. Any further information you receive and cite in defence of your conclusions will be perpetually undermined by the Sceptic. Should this worry anyone?

We've already argued that where a method of comparative evaluation is effective, then we can expect to converge on the truth, but this argument relied on the fact that assessment of a particular

[14] The proof for this result is to be found in the Appendix of [52]

claim (i.e. the grue hypothesis) took place at a relatively high level of abstraction with reference to a broad theoretical setting which admitted empirical test and corroboration. What if our theories are not amenable to such testing procedures? Are there methodological criteria for motivating choice over metaphysical hypothesis?

1.3 Choice Criteria and Reliability

To sum up briefly. We have considered a number of arguments for why we should expect our theory improvement methods to result in a true theory. So scepticism about knowledge in general seems ill advised. But what of the traditionally problematic cases: metaphysics, morality or the number of stars? Surely some scepticism is here warranted. Can we, for instance, undermine the argument that we are likely to converge on the truth, if our choice mechanism is effective? Consider a case where there is no reliability of IBE?

The crucial premise relates to the effectiveness of comparative reasoning. If such reasoning fails, we should not expect any kind of successive increase toward truth. But in Kelly's argument the effectiveness of comparative reasoning was determined by reliability of inference. The reliability of inferences based on metaphysical conjecture or moral intuition cannot be gauged in the same way as we assess the reliability of empirical theories i.e. by means of predicting the actions of observables. Assuming that no *a priori* justification for the use of IBE is forthcoming we have to resort to an *a posteriori* defence of our choice moral or metaphysical intuitions. It is, to my mind, far from obvious that we have no means of assessing the effectiveness of comparative reasoning with respect to metaphysical conjectures or moral precepts. Nevertheless, here's a reason to think comparative reasoning fails in the case of metaphysics.

> The moral derived about IBE and non-deductive arguments generally applies here as well. So long as we are trapped in the domain of inference alone, we have no way whatsoever of determining when the arguments we use are reliable. We have no idea whether we have exceeded their domains of applicability.[15]

Here's a reason to doubt that reason. Assume, as above, that no degree of effectiveness will ever accrue to any metaphysical reasoning, then no consideration of the sub-benefits of any metaphysical theory could motivate us to adopt said theory wholesale. We could not definitively determine whether, for instance, Lewis' account of counterfactuals was better than any other, or vice versa. Mood might take us one way on a given day and another the next. But since, it's uncontested that modal reasoning of the kind Lewis' theory vindicates is indispensable, we have put ourselves in a position where we are forced to admit an inability we can ill afford. Worse still, we could not ever mount a reasoned rejection of Lewis' theory in favour of an alternative, but this is precisely what our assumption sought to motivate. So the assumption undermines its motivation.

What could you say (on metaphysical grounds) to the grue-advocate? On face value the hypothesis is consistent and not immediately false so do we have to tolerate it until it's empirically confirmed or denied? Can we not reason about it on metaphysical grounds? Have we not already been doing so? Metaphysicians or moralists (of any stripe) need not rely on predictive reliability as a criterion for developing theory preference, but they need to be able cite some some criteria. The task before us seems to involve figuring out what kind of criteria are apt for use in any particular domain. Surely Ockham's razor applies here too?

Regardless, the convergence results discussed by Lipton and Kelly relied explicitly on the fact that our choice procedures were aided by empirical corroboration. But as we've seen above not all our theories are amenable to such kinds of corroboration, so what kind of reasoning takes place

[15] [43]pg184

in those domains? One natural suggestion, is to think that where empirical reasoning can rely on objective probability measures, metaphysical or moral reasoning will have to make use of subjective probability measures or perhaps we might cite methodological conventions of metaphysics. Both of these options will be explored in some detail later in this thesis. For the moment we take a more neutral stance and argue simply that the all our above argumentation implicitly relies on a method of theory confirmation called minimalistic confirmational holism; a method in which we seek to exclude the grue-hypothesis on the basis that it cannot fit into our most defensible theories, and since we should adopt our most defensible theories, we will never adopt the grue hypothesis. In particular we shall try to demonstrate that any restriction of the state-space which would rule out grue-like hypotheses can be well motivated by appeal to the in-built restraints which underlie our processes of theory formation and maintainment.

1.3.1 Generalising the Picture

We have been arguing that we can effectively sort between genuine and grue-like theories on the basis that grue-like hypotheses cannot fit well into the overall intellectual mileau which is our theory of the world. In other words, we have stated that underdetermination problems undermine what counts as rational belief. Taken as a explanation of our observations the grue hypothesis is an arguably rational suggestion. But if we hope to maintain any kind of notion of rational belief we need to privilege either the green or grue hypothesis as an explanation for our observation. On this model the primary purpose of an explanation is to induce either belief or knowledge. We cite laws about the properties of emeralds to induce particular kinds of expectations. These laws are, in their turn, motivated by the broader theory in which they are formulated.

We distinguish between justification and explanations as follows: where a justification is used to induce belief in a particular proposition ϕ because certain a set of facts Π have occurred, an explanation is used to induce a belief in the (possibly hypothetical) occurrence of the facts Π because the claim ϕ is believed (or known). Both justifications and explanations are answers to why-questions; although arguably an explanation tracks the inverse of a justification relation. An explanation is factive if it tracks a factive justification relation i.e. if the fact that our future emerald turn out to be green, is sufficiently explained by the fact that all past emeralds were green. Or at least, if we can say truthfully, that the grue-hypothesis was made improbable by the weight of experience, then the future shade of our jewel is explained in some weaker sense by the high probability of the result. This structural similarity between justification and explanation ensures that our procedures of explanation are plagued by many of the same problems as our procedures of justification.

In particular we are subject to problems of underdetermination. We state the problem as follows:

UD For any claim ϕ and any candidate explanatory theory Ψ, there is at least one rival explanatory theory Π such that Π is a contrary of Ψ and both are equally explanatory.

This accounts for why it is typically problematic to specify the relevance of some theory to the available evidence. A corollary of this is the observation that since the explanatory relation is underdetermined generally, it is underdetermined in specific instances too. As such, for any part π or ψ of Π and Ψ respectively, we can claim that said part explains ϕ, or fails to do so just when we are prepared to entwine π or ψ in a further explanatory network. This amendment to our theory allows us to include our theoretical considerations amongst further theoretical considerations. Let $\pi \subset \Pi$, then for some amendment to our theory, we find a theory Δ, such that $\pi \subset \Pi \subset \Delta$ and (π explains ϕ) $\in \Delta$. The only constraint on finding Δ is that it must be such that $\Pi \cup \Delta \nvDash \bot$. Such a theory will always exist since Δ can be arbitrarily large and contain a total translation of all claims in Π so that the truths of Π turn out to be compatible with the claims of Δ after all. In this manner we can find all the explanations we care to. So we have a kind of Quinean observation.[16]

QUD Any theory Π can be reconciled with any recalcitrant evidence ϕ by making suitable amendments to our other assumptions about the explanatory links in our theory.

This is not a trivial problem but it does admit a solution. We need to block the move in which a theorist can simply appeal to a more inclusive explanatory theory. Our explanatory resources are finite, and incessant theory change is no more viable than the flight of pigs. It is an *a priori* fact that we may only consider a finite number of explanatory claims, and the acceptance of some excludes the viability of others, so long as consistency is a concern. Another way of construing this strategy is to see our restriction as a way of blocking ascent to meta-level considerations. This move is then motivated by the observation that most situations call for reasoning from your beliefs, and not from beliefs about your beliefs. As such, patterns of plausible explanations are likely to proceed from establishing that explanatory links such as "π explains ϕ" already exist within our theory Π. Intuitively, we can think of this as a constraint on relevance since we have no way of ensuring that arbitrary extensions of Π will be relevant to the considerations of ϕ.[17] Not only is this kind of constraint importantly motivated as a response to underdetermination arguments, but it is also descriptive of scientific practice.

> Typically, the theory will contain a great many hypotheses, and a given experiment or collection of experiments may fail to measure values of more than one quantity in the theory. To determine the value of one the latter quantities the use of several hypotheses in the theory may be required, and the determination may proceed through the computation of intermediate quantities or combinations of such.[18]

This is the notion of *bootstrap confirmation*[19], where we use parts of the theory to confirm other parts, thereby confirming the whole. The result is that something very like our initial theory is confirmed. This we have been arguing is somewhat inevitable, as we would wish the support structures for our beliefs to rest within the confines of our active theory; this "new procedure, in effect, has us deduce *instances* of laws from singular statements and other laws" within our existing theory.[20] Given our restriction on theory expansion, there will be a finite number of internal expanatory connections we can observe in a theory. The discovery of these connections allows us to simplify the theory. For instance, in book III of the *Principia* Newton uses his first two laws to deduce Kepler's laws which ensure that the centripedal force enacted on a planetary body in our solar system is in inverse proportion to their distance from the sun. Despite the fact that this conclusion is somewhat inexact, because it doesn't factor for the mass of other planetary bodies in the solar system, we do not need to reject this kind of corroboration. We can accept this claim because Newton can argue from independent observational evidence within astronomical theory that the mass of the planets in our solar system is not sufficient to make our calculations woefully wrong.

This strategy of explanation has a claim to be descriptive of scientific practice, and in a certain sense, logically inevitable. Failure to subscribe to such a method of confirmation will lead to undetermination. We argue that while claims are to be corroborated against the background of a theory, no claim is tested against an infinitely expanding background. Furthermore some theories are more plausible relative to first premises than others. This plausibility ranking is not arbitrary as we can cite positive criteria for our particular choices. Hence, underdetermination is empirically false, and the bootstrap theory of confirmation explains it's falsity. We should also note

[16] For details surrounding the history of this argument see [66]pg385-390. But it is also worth consulting Quine directly in [74]

[17] But this restriction can also be motivated independently as an application of the *stalwartness* principle discussed in Kevin Kelly's contribution to [11].

[18] [39]pg409.

[19] Introduced by Glymour in [39]

[20] [39]pg418.

that the structure of the solution accords well with the solution to Goodman's underdetermination problem. Instead of choosing a primitive ranking criterion by which to restrict theory choice, we constrain our theory choice by a voluntary focus on the explanatory power of the facts already available to us. In particular, the information which indicates the patterns of the laws which underlie our projective estimates. The rationale of both the bootstrapping method and our rejection of the grue-hypothesis is similar in structure, and both stand or fall together. Whether we appeal to the implausible consequences of accepting grue-like claims, or the methodological restraints of Ockham's razor. In both cases the connection between theory and evidence is determined by the theorems of our meta-theory i.e our procedural standards both of which are well motivated from within the theory. This can be construed as an inclination towards bootstrapping explanations or particular types of projective estimate. This is as it should be, the logic of the meta-theory is not easily changed, or undermined. Nor should it, as we expect it to be true if anything is. Failure to do so, paves the slippery slope on which the sceptic invites us to slide.

More generally underdetermination problems are to be resolved by observing that the breadth of logical space is not the only domain of rational thought. We can adopt heuristic rules of reasoning motivated by stricter principles than mere deductive possibility. These motivations are what we seek to uncover when faced with an undetermination problem. They are the constraints on rational belief.

1.4 Conclusion

In this chapter we have offered a presentation of a particular riddle about inductive inference. We have sought to demonstrate that for reasons of explanatory implausibility, we are justified in our rejection of the grue-hypothesis, and in particular the false generalisation which underlies it. In effect this chapter has been a case study of explanatory plausibility. We have assessed the plausibility rankings of rival theories and seem some suggestions as to how this ranking may be achieved. The open question is whether we can take the morals learned in this discussion and develop a systematic approach to deriving solutions for abductive problems generally.

In the next chapter we shall look at an explicit attempt to formulate the link between theory and evidence in terms of Bayesian conditionalisation and update. The idea will be to see whether a straight forward account of probabilistic confirmation theory can be used to formalise the notion of theory choice and explanatory priority. We wish to see if the Bayesian setting is such that there is a natural way to rank theories in order of preference, perhaps a preference based on their respective explanatory power so as to motivate rational choice. We will show that it cannot, since the Bayesian model is not suitably fine grained. As such we shall attempt to motivate an approach to reasoning which can factor for a wider species of qualitative and quantitative reasoning based on the adoption of plausible heuristics and probabilistic estimates. Which is to say that, over the course of the next two chapters we shall examine two paradigmatic models of belief and knowledge with a view to capturing the core features of the grue-problem. We shall show that neither model satisfies (a) the necessary condition of being able to articulate the underdetermination problem, and (b) the adequacy condition of being able to address the underdetermination problem.

Chapter 2

Probably True: Gradations of Possibility

> *“Don’t let us forget that the causes of human actions are usually immeasurably more complex and varied than our subsequent explanations of them.”* - Dostoyevsky in *The Idiot*

2.1 Introduction: A Quantitative Model of Belief

What kind of ranking should we expect to hold of our beliefs? If, as seems fair, we admit that belief comes in more degrees than two, we might wish to represent beliefs numerically. There is a reasonable motivation for this move; consider a piece of evidence and a theory where the fact of the evidence renders the theory more probable. We can represent this increase numerically only when we can assess our belief in the theory as being less than true and more than false. Similarly, when choosing between two theories and we might wish to accept the latter over the fomer on grounds of its greater probability, not merely because we know it to be true. In short we need to be able to represent uncertainty.

In this chapter we will elaborate the standard Bayesian approach to belief representation, and point to some problems arising. In particular we will note a categorical failing of the Bayesian system, i.e. the inability to represent heuristic principles and non-probabilistic reasoning, and the role of such principles in explanatory reasoning. We conclude this chapter by using these deficits to motivate another approach towards the representation of belief, which can be used to better capture explanatory reasoning.

We shall present a bare bones axiomatisation of Bayesian theory indebted to Halpern’s discussion in *Reasoning about Uncertainty*.[1] We shall first present the theory largely without comment and then turn to consider the standard Dutch book arguments which are used to motivate the adoption of these axioms.

2.1.1 Bayesian Probability Theory

As in standard probability theory we have a structure $(\Omega, \wp(\Omega), \sigma)$, where we let Ω be our sample space, and σ be our probability assignment on the powerset of our statespace Ω. That is, σ is map from $\wp\Omega$ into the set of real numbers between 0 and 1.

Probability function $\sigma : \wp(\Omega) \mapsto [0, 1]$

[1] See [45]

Subsets of Ω are typically called events, or analogously propositions describing those events. However, given that we wish to use this probability metric to represent an agent's subjective estimates of probability, we should treat the subsets of $\wp(\Omega)$ as a set of propositions which σ maps to probability values representative of an agent's credences. If the idea is that an agent would represent beliefs to themselves probabilistically we should be able to associate certain propositions with a particular numerical value determined by the probability function σ. In this case the probability determines an agent's degrees of belief. For the moment we ignore whether these beliefs are motivated by an objective or subjective probability measure. Some questions arise. For instance, if an agent i believes the proposition p to some degree n, what kind of threshold probability should distinguish belief from unbelief? If σ: p $\mapsto$ .7, do we observe a sufficient strength of belief to indicate i's acceptance of p? Is there a non-arbitrary point? This prompts a more general question; what kind of constraints should we impose on σ.

Note that Ω is a sample space closed under the operations of union and complementation, which is just to say that if E, H and T are in Ω, then E $\cup$ T, H $\cup$ E.... and $\overline{E}$, $\overline{T}$ and $\overline{H}$ are in Ω. This makes sense on the assumption that you might wish to assign probabilities to conjunctive propositions, and negative claims. Now consider the following constraints.

(P1) $\sigma(\Omega) = 1$

(P2) $\sigma(\mathrm{E} \cup \mathrm{T}) = \sigma(\mathrm{E}) + \sigma(\mathrm{T})$, if E and T are disjoint sets subsets of Ω.

It's straightforward to see that the above constraints entail that $\sigma(\emptyset) = 0$, and more importantly the probability measure satisfies the property of finite additivity, which is just to say that the generalised version of (P2) holds.[2] But why would we want to insist that the probability of the union of a set is equivalent to the sum of it's parts?

Before answering this question, we must first observe that we can calculate the conditional probability of an proposition T given E where $\sigma(\mathrm{E}) > 0$.

CondP:

$$\sigma(T \mid E) := \frac{\sigma(T \cap E)}{\sigma(E)}$$

Although strictly speaking we might better think of conditional probability space as an extension of standard probability space in which we have:

Definition (Conditional Probability Space) CPS = (Ω, $\wp(\Omega)$, Ω', σ) where Ω is as before but σ is a function $\sigma : \wp(\Omega) \times \Omega' \mapsto [0, 1]$ and Ω' is non-empty subset of Ω closed under boolean operations.

The idea is that we take conditional probability as our primitive so that every probabilistic statement comes in the form $\sigma(\mathrm{T} \mid \mathrm{E})$, and we derive particular values from our conditional information. But to do this we need, for every statement in our language, to conditionalise of some subset of the statespace. That is to say if every statement is of a conditional format, then every claim involves the assumption that some element of the statespace is given. We use Ω' to indicate the region of the statespace we take as given. On this understanding the natural analogues of (P1) and (P2) come out as follows:

(CP1) $\sigma(\mathrm{E}|\ \mathrm{E}) = 1$ if E $\in \Omega$'

(CP2) $\sigma(\mathrm{T}_1 \cup \mathrm{T}_2\ |\mathrm{E}) = \sigma(\mathrm{T}_1\ |\mathrm{E}) + \sigma(\mathrm{T}_2\ |\mathrm{E})$ if $\mathrm{T}_1 \cap \mathrm{T}_2 = \emptyset$ and T_1, $\mathrm{T}_2 \in \wp(\Omega)$ and E $\in \Omega$'.

[2] i.e. that $\sigma(\mathrm{E}_1) \cup ....(E_n)) = \sigma(\mathrm{E}_1 + ...\sigma(\mathrm{E}_n))$. Halpern reports that the proof is an easy induction in [45]pg16.

(CP3) $\sigma(\mathrm{T} \cap \mathrm{E} \mid \mathrm{H}) = \sigma(\mathrm{T} \mid \mathrm{E} \cap \mathrm{H}) \times \sigma(\mathrm{E} \mid \mathrm{H})$ if $\mathrm{E} \cap \mathrm{H} \in \Omega$' and $\mathrm{T} \in \wp(\Omega)$.

Halpern analyses the third axiom, and shows it to cover some natural properties of the conditional probabilistic estimates; such as the constraint that when conditioning on E, everything should be relativised to E.[3].

The general idea is that we can define the notion of a conditional bet $\sigma(\mathrm{T}|\ \mathrm{E})$ as being tantamount to the claim that if E happens then T is credible to the degree determined by σ. The axioms are justified because if they are not assumed irrational behaviour is assured. We proceed as follows: we assume, for instance, that the probability calculus does not satisfy the additivity condition and the Bayesian agent remains rational. We elaborate a Dutch book argument to prove that the Bayesian cannot be rational under such conditions. This is an argument to a contradiction where rationality is defined in terms of pursuing non-negative betting outcomes. It seeks to show that the probability axioms are required for a rational account of probabilistic reasoning.

Defence of the (Basic) Axioms

Allow that our degrees of belief are expressed by the degree of fiscal commitment we are prepared to make while betting on the corresponding proposition. So we distinguish between the degrees of belief determined by σ which determine our willingness to bet and the stakes α which determine how much we will pay for a particular bet. There is a transitive preference order over bets such that we prefer the bets that give more money for lesser stakes. The preference order is sure to comparatively rank a bet and its complement. Let (P, α) is the bet such that if P occurs then you win α and the bet costs $\sigma(\mathrm{P})\alpha$, and if it doesn't occur then you lose where α is thought of as your stake in the bet. A bet is deemed acceptable if an agent takes the probability of $\sigma(\mathrm{P})$ to be sufficiently high to outweigh the risk induced by the stake required to bet on (P) i.e. an agent buys a fair bet (P) for cost of at most $\sigma(\mathrm{P})\alpha$. So imagine that $\sigma(\mathrm{P}) = 1/2$ and you stand to gain £1, then you will be willing to pay at most £.50. So your winnings will be £1.00 - £0.50 as expected. The synchronous Dutch book argument states that if the axioms of probability theory do not hold, then there is a series of bets that an agent will find simultaneously acceptable, but which ensure a net loss.

Proof: To show (P1) assume $\sigma(\Omega) \neq 1$. There are two cases, we show one. Let $\sigma(\Omega) > 1$. Hence 1-$\sigma(\Omega) < 0$. Then take a bet such that the stake $\alpha > 0$, so given that (Ω) is always true we calculate our winnings as the stake minus the cost the cost determined by our belief i.e α- $\sigma(\Omega)\alpha$ = (1 - $\sigma(\Omega))\alpha < 0$. So any single bet ($\Omega$, α, $\sigma(\Omega)$) with $\alpha > 0$ will ensure a loss, since the pay out will always be α-$\sigma(\Omega)\alpha$ regardless of the values, because Ω is always true. This is a negative quantity given our assumption.

To show (P2) assume $\sigma(\mathrm{H} \cup \mathrm{E}) \neq \sigma(\mathrm{H}) + \sigma(\mathrm{E})$ where H and E are mutually exclusive. Notation: let $\sigma(\mathrm{H} \cup \mathrm{E}) = \mathrm{q}_1$ and $\sigma(\mathrm{H}) = \mathrm{q}_2$ and $\sigma(\mathrm{E}) = \mathrm{q}_3$. Again we show one direction. Let $\mathrm{q}_1 < \mathrm{q}_2$+$\mathrm{q}_3$. We fix the stake on each bet as follows Bet 1 = St(H $\cup$ E) = -α, and Bet2 = St(H) = Bet 3 =St(E) = α. So we have the following payoff table.

[3] cf. [45]pg75 for details. The broad idea is that it should be possible to calculate the proportion of educated theorists in any room, by multiplying the fraction of theorists by the fraction of educated people in the room. Note that we can recover the unconditional probability measure σ for $\sigma(\mathrm{E})$ by defining a translation $\sigma(\mathrm{E}) = \sigma(\mathrm{E}|\ \Omega)$.

Table 2.1: The Payoffs

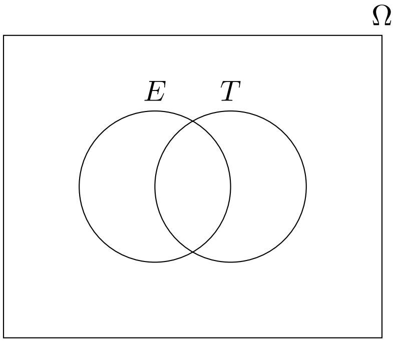
H	E	H ∪ E	Bet1(H∪E)	Bet2(H)	Bet3(E)
T	T	T	$-(\alpha$-$q_1\alpha)$	α-$q_2\alpha$	α-$q_3\alpha$
T	F	T	$-(\alpha$-$q_1\alpha)$	α-$q_2\alpha$	-$q_3\alpha$
F	T	T	$-(\alpha$-$q_1\alpha)$	-$q_2\alpha$	α-$q_3\alpha$
F	F	F	$q_1\alpha$	-$q_2\alpha$	-$q_3\alpha$

By assumption (H) and (E) are mutually exclusive, so we can effectively ignore the top line of our table. If all events fail to occur then you straightforwardly lose you money to bookie ensuring a loss. Assume Bet2 is true and Bet3 is false, then the payoffs for the set of bets is (α-$q_2\alpha$-$q_3\alpha$-(α-$q_1\alpha$)) = α(q_1-(q_2+q_3)) < 0. Hence, so long as $\alpha > 0$, we are ensured a net loss. ⊣

More intuitively this is a Dutchbook-able payoff distribution since the bookie can offer to sell the individual bets with potential winnings of α, but himself buys the disjunctive bet from you for the same potential winnings at fair prices. This ensures you a net loss on any of the payoff distributions since the amount of money you place on each bet plus the payout you need to make will exceed any winnings you might accrue.[4] So it can't be rational to think the degrees of belief don't obey the probability calculus. In addition, Van Fraassen offers a diachronic argument (discussed below) which shows us the importance for conditionalisation and its role in planned betting. The idea is that if we violate the axioms of conditional probability a Dutch book can be formed, and the probability of achieving any gain in such a circumstance is 0, we should be able to avoid entering into such gambles. That is to say that belief change should occur only under conditions which factor for the consequences of those changes i.e. when we conditionalise. This latter claim assumes an idealised kind of a reasoner, who can (and does) constantly calculate the odds of gain that emerge from future actions.[5]

Bayesian Conditionalisation and Update

Recall our definition of **CondP**, you might want to ask why this procedure should hold. Curd and Cover offer the following rationale:



> Suppose that you are told that a dart has been thrown, randomly, at the figure and has landed somewhere inside the T-circle. Given that the dart is inside the T-circle, what is σ(E|T), the probability that the dart is also in the E-circle. The answer is simple: it is the area common to both circles divided by the area of the T-circle. In other words given that probabilities are proportional to areas, σ(E| T) is equal to σ(E ∩ T) divided by σ(T).[6]

[4] A very similar proof is used to show that violations of countable additivity also result in a Dutch book situation.

[5] These arguments are originally owed to Ramsey and de Finetti. A discussion appears in ”Why Conditionalise” in [62]pg403, and other (more formal) justifications for using probability to represent reasoning under uncertainty can be found in [45]pg77 -78 and [27]. The intuitive idea behind all Dutch book arguments is the same in each setting.

[6] [66]pg630.

From this observation, it becomes quite clear that the conditionalisation operation is as it should be, but can we use this to assess issues of theory choice? Surely we should be able to assess the probability of a theory given some evidence. First we simplify things slightly. By **CondP** we derive the following rule since σ(E) cancels out.

$$\sigma(E \cap T) = \sigma(T \mid E) * \sigma(E).$$

This latter rule is called *the general multiplication rule.* It allows us to assess the probability σ(E ∩ T) in two stages, arguably this is easier than computing the probability directly. From this simplification we draw (finally) a statement of Bayes' theorem by noting the equivalence of σ(T ∩ E) = σ(E ∩ T), and substituting the former into the multiplication rule, and the latter into CondP. We get:

$$\sigma(T \mid E) = \frac{\sigma(E \mid T) * \sigma(T)}{\sigma(E)}$$

This result is uncontroversial (even elementary) in probability theory, but given our intended application we should specify what Bayes' theorem means. First we distinguish the parts: σ(T) is the *prior probability* of the theory T, while σ(E) is the prior probability we ascribe to the evidence E, this can be thought of as the *expectedness of* E. Furthermore the probability σ(E ∩ T) is the expectedness that our theory and evidence coincide. Bayes' theorem states that the probability of a theory T given evidence E, is equal to the chance that the truth of our evidence and theory coincide, divided by the chance our evidence is true. Now consider how we develop a Dutch book style argument for conditional probabilities.

Proof Assume that σ(E) = $q_1 > 0$. If σ(T | E) = $q_3 \neq \sigma$(T ∩ E) ÷ σ(E) = q_2/q_1 i.e. contrary to the definition of **CondP**, then we can show that there is a Dutch book. Again the proof is in two parts, we show one direction i.e. that $q_3 > q_2/q_1$ is contrary to reason. Let's set the stakes as follows: Bet1 = St(T ∩ E) = -1 and Bet2 = St(T | E) = 1 and Bet3 = St(E) = q_3.

Table 2.2: The Payoffs

H	E	H ∩ E	Bet1(T ∩ E)	Bet2(T | E)	Bet3(E)
T	T	T	$-1+q_2$	$1-q_3$	$q_3-q_1q_3$
T	F	F	(q_2)	0	$-q_1q_3$
F	T	F	(q_2)	$-q_3$	$q_3-q_1q_3$
F	F	F	(q_2)	0	$-q_1q_3$

On this table we have input the relevant values for α.The payoffs for the conditional bets are determined in such a way the bet fails just when E is false. The values allow us to determine the payoffs for this set of bets. As such we can see that for any values of H and E, the value of the set of bets will be q_2 - q_1q_3, so by our assumption we know that $q_2 < q_1q_3$ and this quantity will be negative for any truth-values. ⊣

The Diachronic Dutchbook

Perhaps more interestingly we can think of the creation of a diachronic Dutch book where the bets are time indexed as follows: (1) the bookie offers to sell the unconditional wager whether (E) at cost σ(E) = σ(T) = 1/5, where the bet pays £.10, and buy the conditional wager (T | E) at cost σ(T | E) = 1/2 which pays out £1.00. Then (2) we have two choices either E turns out to be true, or turns out false. If E is false, then the conditional wager is called off and no one wins or loses, so you have just bought the unconditional wager and lost. If E turns out true, things are slightly more complicated - you should increase you belief in the probability of T, above your belief in the conditional probability. Suppose σ(T) = .6, after you learn E. The bookie now offers to sell you the unconditional wager (T) at a fair price determined by your conditional belief, which pays out £1.00. So either (T) turns out true (in which case you pay the bookie for his bet) or false (in which case you lose your most recent bet), in either case your expenditure has exceeded your gain.[7]

The point of Van Fraassen's Dutch book argument[8] is that we should only accept a particular style of belief-update based on our conditionalised expectations. Observe that the conditional method of update is sure to catapult new propositions into a range of credence. Whatever the subjective threshold for belief, the Bayesian model automates belief change procedures, and belief change/tesiting must occur constantly on pain of the diachronic Dutch book argument. Since the diachronic Dutch book argument proceeds on the basis of our ignorance about our own future credences. We might think to amend the nature of Bayesian update to respect this temporal parameter by adding an extra constraint on our probability function σ.

(Refl) $\sigma_t(E \mid \sigma_{t+x}(E) = i, 0 \leq i \leq 1) = i$[9]

Van Fraassen calls this the Reflection principle, and observes that it is warranted by the rational coherence expected of an agent by others. If you make an assertion or commit to the likely state of your future belief, you implicitly assume that you have evidence for the future belief, but in that case you ought to believe now, what you think you will in the future. This is just to say that your conditional belief $\sigma_{t+x}(\mathrm{H} \mid \mathrm{E})$ depends on the value $\sigma(\mathrm{E})$ which you have either (a) already assumed at time t, or (b) you assume knowledge of $\sigma_{t+x}(\mathrm{E})$ which is yet to be determined at the earlier time. Both moves are fallacious, the latter presumes omniscience and the former is an illegitimate presumption of the uniformity of belief. Evidently, there is a clash between two ideas: (1) if the beliefs of an agent can be fully determined by their priors, then we should not expect any confessions of ignorance from an agent reflective enough to predict their own future beliefs, on pain of Dutch-book arguments. But (2) we do tolerate such statements of ignorance. The moral, Van Fraassen suggests, is that the probability assignment σ is adopted voluntarily, and not simply derived from some primitive (and total) assignment function. We opt into the Bayesian belief estimates just when on reflection we are prepared to defend those estimates. This perhaps explains the queasiness we have with the idea that each and every proposition should come with a probability assignment; we should only expect those propositions voluntarily endorsed to some degree to meet the standards of rationality associated with the probability calculus. This is all well and good, but what does the Bayesian rationality have to do with theory assessment?

Subjective Vs Objective Probabilities

On the Bayesian setting it is standardly assumed that the probabilities reasoners factor for are subjectively determined estimates. With van Fraassen we have pointed to some reasons to doubt that our subjective probability estimates define a total function. Perhaps this motivates the idea that some of our probabilistic estimates, are more the record of objective probabilities than the history of guesswork. The objective probability measures is distinguished from the subjective measure on the basis that objective probabilities are assigned only to those probabilistic systems for which have a record of the patterns of that system. So for instance, the history of coin flipping attests to equal odds for heads or tails. Whereas a subjective interpretation of a coin-flipping event, could allocate a higher, whimsical, preference for heads, perhaps on the presumed basis of a false coin. In short, objective probability values, are "read off" from the record, while subjectivist probability values are "imposed" from without for reasons often known to none. In the next section we will consider what is the appropriate interpretation of the σ-function.

[7] This argument is owed to David Lewis via Teller see [62]

[8] cf. "Belief and the Will" in [1]

[9] The subscripts t and $t{+}x$ denote positions in time at which probabilities are assigned.

2.2 Bayesian Confirmation Theory

Returning to the issue at hand; how might Bayesian theory help with issues of theory choice? Wesley Salmon[10] derives the following rule to help ajudicate in cases of theory choice, where we have a number of theories $\mathrm{T}_1....\mathrm{T}_k$ that are mutually exclusive, exhaustive of the state-space. Ultimately we would like to see if such a theory preference mechanism is able to help resolve undeterdetermination problems.

Salmon's First Proposal

$$\sigma(T_i \mid B \cap E) = \frac{\sigma(T_i \mid B) * \sigma(E \mid T_i \cap B)}{\sum_{j=1}^{k}[\sigma(T_j \mid B) * \sigma(E \mid T_j \cap B)]}$$

In the above formula every theory is assessed conditionally on the basis of the background information and the current evidence. This makes even reasonably simple calculations slightly cumbersome, but it's a fair approximation of the type of considerations which should be factored for in theory choice. The former observation stands as a criticism of the Bayesian method in so far as it means the probabilistic calculus cannot account for the large part of our everyday reasoning. However, accepting Salmon's proposal, whether we treat the function σ as an objective frequentist measure or a subjectivist estimate elicited at request, we have a method for determining a normative standard for belief in some given theory T_i. Yet, a Bayesian account allows for the fact that the normative standard is simply that of subjective self-consistency, so different agents can arrive at distinct estimates if we allow their priors to change. This suggests that rationality and objectivity can diverge. Let's look a bit deeper.

The Parameters

Simply by the limited ability of human beings to survey large sample spaces we should not expect their subjective probability estimates to line up exactly with the objective probabilities. However, given the structure of the scientific enterprise and the division of labour we might think that scientists should be in a position to trust the reports of other scientists. In this manner, each reported test of a given hypothesis could be seen to impact the priors of any agent. Where we observe that some evidence E is relevant to the assessment of a hypothesis H, we can inaugurate the following standard.

$$\text{E confirms H} \leftrightarrow \sigma(H \mid E) > \sigma(H)$$

We then allow that the scientific reports which seem to confirm H, are to be treated as a predictor of an agent's priors just when an agent is linked in a network through which those reports are filtered. Assuming some kind of procedure can be developed for systematically integrating the reports of others, we might want to think that our priors track objective results. However, think of the grue-case, and allow that E is the observed history of emeralds, and H is the grue hypothesis. Why suppose that σ(H) is an objective estimate? It's consistent with our objective results but that doesn't quite cut it. By the structure of the grue problem E equally confirms H and ¬H. If we have accepted some cases of unobservably grue-emeralds in our evidence i.e. assigned an arbitrarily high value to σ(H), we could conclude that the grue-hypothesis was true, but this seems to just beg the question of how we have determined our evidence to be relevant to the assessment of H or ¬H, when conditionalising makes no difference! In particular, why these emeralds are grue, and not just green?

This is far from clear in the case of our estimations of the expectedness of evidence since our expectations of E should (at least implicitly) depend on our understanding of the theory T, and

[10] cf. "Rationality and Objectivity in Science" in [66]pg555

importantly, its interactions with the background B. But it's far from obvious that such connections are simply a matter of deductive consequence, and even if they were, none of us are logically omniscient. If only probabilistic connections are appropriate our problem repeats and we find ourselves in a regress.[11] It's even more obscure if we seek to evaluate σ(E) without factoring our background knowledge. In general the relevance of our background beliefs to the expectedness of evidence σ(E) cannot be calculated *a priori*. Qualitative considerations which respect the context of assessment need to be included in the estimation of σ(E). In any case it's very difficult to see why our expectation of E should be accurate when based purely on subjective estimates.

Preliminary Doubts

Worse still, there seems no good reason to believe our sample space is accurately carved up by the mutually exclusive theories $T_1....T_k$. The assumption that our sample space Ω is totally partitioned by our carving is too generous, for the natural reading of the claim is that all scientific theories to be considered, are all those that will ever be discovered. This is evidently false, and clearly a formal analogue of the inductive projection fallacy noted by Hume. This ensures that **Salmon's First Proposal** is too heavily idealised.[12]

A similar criticism can be deployed against the stable estimation or "washing out" theorem of Leonard Savage. The idea is that we can defend Bayesian theories as appropriate for scientific reasoning because under certain specifications, successive operations of conditionalisation on increasing information will ensure that the Bayesian's information converges to the truth. We recount a somewhat simplified picture of this convergence thesis as follows: suppose we have an urn with 80 balls in it, at least one of which is black. The idea is that if we randomly sample (without replacement!) from the urn for a sufficiently large number of times we will be able to rule out the false hypotheses of the form "n balls are black". Glymour criticises this view on two fronts: (1) it's not at all compelling for non-Bayesians to observe that Bayesians will get it right eventually - even broken clocks are correct twice a day but that's small consolation if you miss the bus most of the time. (2) the conditions required to derive Savage's result depends on too many idealisations not seen in actual practice. For one, we have to be an a position to distinguish each of the mutually exclusive hypotheses. This rarely happens. Our evidence can be collected by way of random sampling, where we expect each evidentiary observation to be independent of all others. This never happens. The last point is crucial. Cumulative evidence gathering in science tends to increase a bias, if the sequence consistently corroborates some particular hypothesis. Hence this defence of Bayesian rationality reinforces the view that the setting is best reserved for idealistic models of normative reasoning.[13] Savage's theorem does not convince us that our best reasoning will assure our arrival at the truth on any practical matter.

Comparative Evaluation

As we observed above there is a slight mystery over how we specify the value of some of the parameters which go into the assessment of our belief. Particularly our evaluations of the connections between theory and evidence. In addition we noted a problem with **Salmon's First Proposal**. We now consider an amendment to the last suggestion which aims to provide a *Bayesian algorithm for theory preference* where we compare the probability of two theories directly factoring for identical background.

[11] Assuming a background B_i is always factored at each preceding stage of theory construction.

[12] We do not wish to suggest that Salmon was unaware of this deficit, rather the opposite. He uses exactly this observation to motivate his second proposal discussed below.

[13] For further discussion cf. [66], in particular Clark Glymour's "Why I am not a Bayesian". A more compelling theorem about convergence to the truth in the limit is proved in Kelly's [53]pg228, although this result is not directly reliant on Bayesian assumptions. For a further discussion about the idealisations required to achieve convergence under Bayesian assumptions see [53]pg330-337.

Salmon's Second Proposal

$$\frac{\sigma(T_1 \mid E \cap B)}{\sigma(T_2 \mid E \cap B)} = \frac{\sigma(T_1 \mid B) * \sigma(E \mid T_1 \cap B)}{\sigma(T_2 \mid B) * \sigma(E \mid T_2 \cap B)}$$

The benefit of this algorithm is that it avoids the requirement for factoring a complete specification of state space Ω, because we effectively ignore any theories other than the two being considered. This ignores problem rather than solves it. For to get the point where we are deciding amongst two theories we must have performed a prior calculation ruling out all others, or at least be prepared to embark on a (potentially infinite) series of comparisons relying on the transitivity of preference. For reasons stated above, we cannot hope to have fairly dismissed anything more than a sample of our state space. But such a sample is only fairly called a representative sample by begging the question, unless we stipulate that σ is a total and objective probability function on Ω. However, then the utility of the Bayes algorithm is less useful for general issues of theory choice, since there are very few situations in which we have a full objective map of the probability distribution over all possible theories or their components.

It is similarly unclear that we can overcome the problem of how to determine the connection between the theory and the evidence e.g. why is E relevant to the assessment of T but not T'? Evidently such connections do exist, but the Bayesian system cannot factor them. We define confirmational relevance in a probabilistic setting, if the probability of T is raised by the condition of E. However, such a criterion is too coarse grained to distinguish between the grue-theory and the natural alternative, since both theories have their credibility raised equally by all current evidence. Worse still, given the fact that expectations can differ wildly, even if we allow that the structure of the scientific enterprise ensures that some opinions σ_i(X) will converge, we cannot hope that convergence between all probability assignments σ_i amongst a set of agents will occur just when X is true.[14] Intuitively, a set of agents can come to differ in their estimates of the posterior probabilities just so long as their initial estimates are independent[15] of each other. If you grew up in a grue-culture, your prior-estimate of the grue-hypothesis would be reasonably higher than my estimate of the same proposition. But the Bayesian does not provide us the tools to adjudicate between such contrary beliefs.

Significant Objections

Think about how we would try to factor for meta-level considerations of theoretical viability. How does, for instance, the Bayesian account for theoretical simplicity? Here we elaborate an argument which purports to show that Bayesian accounts of simplicity are simply question-begging. Suppose, we assess two theories where we know that T$_1$ is objectively "simpler" than T$_2$. How can the Bayesian account of rationality account for this without imposing specific restrictions on the calculation of the posterior probabilities that account for this very supposition?[16]

First observe that if there needs to be a choice between the two theories, then σ(T$_1$) ≈ σ(T$_2$)[17] and for both theories the expectedness of the evidence σ(E) ≈ 1, otherwise there would no need to make a choice based on the simplicity criterion. More importantly if one theory i.e. T$_2$ is more complex than the next we must suppose that there is a free parameter in the more complex theory such that the objective chance of E occuring is high given σ(T$_2$), and for any alternative assignment of belief values the estimation of E approaches 0. This is supposed to explain why the choice between the two theories is vital - there are no other plausible theories in the vicinity. The notion of objective chance differs from the subjective probability estimate. But factoring for

[14] For more on this point cf. Glymour's discussion of Savage's theorem in "Why I am Not a Bayesian" and the discussion in [66]pg648-9.

[15] We say that a theory T is independent of evidence E, just when σ(T | E) = σ(T).

[16] For convenience we omit direct considerations of a background B.

[17] i.e. the theory are believed to approximately the same degree.

objective chance in a Bayesian setting amounts to the same idea as if an agent were to take a guess about who would win a fair lottery, while knowing exactly how many people would play. So where we have a free parameter in a given theory, we estimate the subjective belief that a particular "player" will win a fair lottery as $\sigma(\text{Player}_i \mid T) \approx \frac{1}{n}$, where n is the number of players in the lottery. A complex theory might have a free variable, but for it to be a theory, there has to be some restraint on the possible values of this parameter. The effect of this parameter is such that when calculating the ratio of the two theories

$$\frac{\sigma(T_1 \mid E)}{\sigma(T_2 \mid E)} = \frac{\sigma(T_1)}{\sigma(T_2)} * \frac{\sigma(E \mid T_1)}{\sigma(E \mid T_2)}$$

We are assured that the prior probability converges on 1, but then the so called Bayes factor, i.e. the second quotient on the right hand side is the vital component. But then applying the total probability rule, we get:

$$\frac{\sigma(T_1 \mid E)}{\sigma(T_2 \mid E)} = \frac{\sigma(T_1)}{\sigma(T_2)} * \frac{\sigma(E \mid T_1)}{\sum^{Var} \sigma(E \mid Var_i)\sigma(T_2 \mid Var_i)}$$

Putting all the values through this equation, we ensure that a strong advantage accrues to the simpler theory because the values for the unfixed parameter bias our eventual theory choice.

$$\frac{\sigma(T_1 \mid E)}{\sigma(T_2 \mid E)} = n$$

So the "simpler" theory is more probable than the complex one because it is less "adjustable". Despite the fact that we've agreed that both theories predict E equally well, the ratio assessment ensures that T_2 is dis-confirmed when compared to T_1. When we factor for a continuous distribution of objective chances, the argument is similar.[18]. This result is a formalisation of a choice method but it was entirely dependent on specifying a question begging probability distribution of the priors contained in both theories. As such the Bayesians cannot claim to have analysed simplicity, but rather they simply provide a representation of it. This is a defect since a normative theory of reasoning should be able able to endorse Ockham's razor.

Another problem emerges when we consider the case of novel information. How might we rate the value of old information compared to the new. Recall that evidence E confirms a theory T, just when conditionalising on E increases the probability of the theory T. But if we account for evidence E_{old} at a time when it is definitively established, we should treat it as "known" or "accepted as true" hence $\sigma(E_{old}) = 1$, but in that case $\sigma(E_{old} \mid T)$ must also be 1. As such, we cannot use E_{old} to confirm T because it will turn out that $\sigma(T \mid E_{old})$ is equal to the value of our prior $\sigma(T)$, which is no advance on what we already knew. This problem suggests another. Let's assume we can distinguish evidence-from-observation from evidence derived from theoretical concerns. Presuming we allot observational evidence more credibility than non-observational evidence, what explains the idea that we take a theory to be more credible than any observable evidence? If the Bayesian appeals to criteria like explanatory coherence or predictive power, they might be able to motivate this discrepancy. However, if we argue that explanatory coherence is not alone sufficient to motivate a high credence, then the Bayesians' credence in the theory T is unwarranted whenever it exceeds their credence in the observable evidence E. In short, there is no good reason to believe a theory more than you believe your observations, and the Bayesian conditionalisation method cannot account for this fine grained property of evidence. The weighted relevance of different species of evidence for a particular theory is impossible to calculate in a Bayesian system. This is a defect

[18] cf. Section 4 of "Simplicity, Truth and Probability in [11]

since it is inadequate to capture the reasonable standards of actual empirical reasoning.

These brief notes are not decisive, but we do think they are indicative. Proposals which seek to utilise the Bayesian model as an exact copy of our inferential procedures leave too many variables unexplained. Worse, the variables they do factor for seem too idealised since the model of Bayesian reasoning lacks any psychological plausibility as a general model of how we reason.[19] In contrast, the Bayesian model does appear to be an excellent model of how we should reason about probability once the degrees of our belief have been resolved. But given the mystery surrounding the workings of the σ-function, we might think that we should only trust this normative standard of reasoning when we can supply objective frequency results[20] for each of the parameters. However, there is a problem here too.

Odds on Paradox

Here we elaborate an argument against the notion that objective interpretation of the σ-function actually helps the Bayesian. In fact we show that neither interpretation is feasible, but since you might think that subscribing only to objective probabilistic information is more likely to ensure consistency and reasonable inference, we show that this is false.

Imagine you're invited to take part in a lottery, your friend (the judge) has assured you it's a fair lottery with 100 tickets. Nepotism hasn't gotten you anywhere, and since you happen to know the judge best, you're convinced that no one else has a way of cheating either. Now suppose that you think you're very rational, and therefore only believe propositions which have been accorded an arbitrarily high probability. Then since you know the lottery is fair, σ ensures that for any particular ticket t_i:

$$\sigma(\mathrm{Wins}(\mathrm{t}_i)) = \frac{1}{100}.$$

Since 1/100 probability is always less than your belief threshold, you come to believe $\neg(\mathrm{Wins}(\mathrm{t}_i))$, and then, by conjunction, you come to believe that no ticket wins. But then the lottery isn't fair. Contrary to our initial assumption.

The conclusion that the Lottery paradox seems to recommend is that either (i) there is no non-*ad hoc* candidate for a threshold for rational belief on a scale of degrees of belief or (ii) your friend is involved in a conspiracy. The argument[21] is that for any candidate threshold *th* which marks rational belief (where we assume the rational belief that the lottery is fair), this threshold is apt to prompt a contradiction since it is rational to be believe, (a) that there is a winning ticket and (b) that for all $\mathrm{tickets}_i$ it is rational to believe that ticket_i is not the winning ticket. By conjunction we get $(\forall \mathrm{tickets}_i(\neg \mathrm{Wins}(\mathrm{t}_i) \wedge (\exists \mathrm{t}_i(\mathrm{Wins}(\mathrm{t}_i)))$. This is contradictory. More worryingly, such a result seems to motivate the conclusion that the entire approach of degrees to belief is somewhat meaningless given that we are forced towards the conclusion that rational belief is marked at the boundary $1 \vee 0$. This is nothing short of the absurdity that rational belief is to be identified with certainty. To bet on the degrees of belief account of rationality, would render you bankrupt. This prompts revenge-stlye arguments over whether it is rational to believe the theory of rational belief characterised by degrees of believe.

[19] This point is best brought home in two ways (1) by van Fraassen style concerns about the bizarre automated style of reasoning required to satisfy the constraints of Bayesian rationality and (2) the now famous experiments by Kahneman and Tversky, e.g. the conjunction fallacy discussed in "Extensional versus intuitive reasoning" in [1]pg110 - 135.

[20] Or at least we should be in a position where we can rule out the risk of small sample bias.

[21] cf [31] for details

The paradox of the preface presents a structurally analogous argument. It asks that we check the consistency of an author who prefaces his newly minted book with the acknowledgement that there is a some fault in the book. But since he has sent the book to the publishers he clearly endorses each particular sentence.[22] So again, by conjunction, he believes every line in the book to be true, and admits at least some falsehoods. Both paradoxes begin with an overriding intuition that there is at least one amongst a set of propositions that is true(false) which we believe[23]. The problem emerges because when we consider each proposition individually we cannot reasonably (dis)believe any of them, and so by conjunction we generate a contradiction with the former observation. Both seem to problematise quantitive accounts of rational belief since the testing of each proposition in the set shows that we actually have a fairly liberal account of rational belief in so far as they show us to be largely unreflective in the numeric self-ascription of rational beliefs. Perhaps, it's better to say, these arguments show that if there is any sound notion of rational belief, it involves a qualitative notion of belief.

To see this we should consider the difference between the two paradoxes. First note that the issue in the Lottery paradox seems to be over the application of the additivity axiom governing σ. However, if we scrap the idea that belief is motivated by a particular threshold of credence, then the whole paradox dissolves where the agent fails to believe the conjunctive claim voluntarily. In the lottery paradox the threshold condition for belief is inappropriate, but intuitively its application is less controversial in the paradox of the preface, since the paradox only gets off the ground where beliefs are already ascribed.[24] Given that we cannot dissolve the paradox of the preface we should commit to the idea that there is a rationale for beliefs we don't know to be true i.e. we can justify the claim that we believe every proposition in our book, without the requisite checking of each, by loose appeals to the coherence of (for example) a chapter within the book. This indicates that the author may withhold belief in particular cases, while nevertheless accepting the whole as correct with qualification. Both paradoxes ultimately suggest that we cannot take the σ-measure of belief to be a decisive, or indeed, tolerable characterisation of belief.

2.3 Problems and Prospects

In this section we state perhaps some fundamental problem with the naive Bayesian setting. Crucially this demonstrates how pressing the need for link between theory and evidence really is. That is to say, we develop a picture which mandates the incorporation of heuristic justifications and explanations in a probabilistic setting, as a meaningful aid in decision problems. In particular we need to be able to factor for multiple notions of belief generation and the basic Bayesian setting does not allow for this. We then conclude the chapter.

2.3.1 Dostoyevsky's Challenge

So you might think that the lottery paradox emerges because our player is inattentive to the relationships between particular probability assignments and their logical consequences. Casting this as a simple observation of inconsistent reasoning, is too quick. To use a fashionable term, we would rather say that there are contrary beliefs but no cognitive dissonance. How might we represent this distinction in terms of a probability calculus? Or why even tolerate the notion?

We should tolerate the notion of cognitive dissonance because it's empirically correct. It's clearly factually descriptive if politician's endorsements are taken as representative of their belief, but a more interesting example is that of the literary character of the Narrator in Dostoyevsky's *Notes From the Underground.* The Narrator believes himself to be sick, and ignorant of any reason

[22] We hope.

[23] i.e. "There is a winning ticket in the lottery", and "There is a mistake in the book"

[24] Essentially the same the point is made by Stalnaker in [1]pg276-277.

why, educated and superstitious, spiteful and unembittered. In short, a man who is ultimately able to find solace in a toothache. For our narrator, the apparent oscillation of moods is nothing less than the observance of his own internal contradictions. Contradictions prompted by the general condition of man:

> [M]an everywhere and at all times whoever he may be, has preferred to act as he chooses and not in the least as his reason and advantage dictated[25]

As it happens, I'm not even sure our Narrator overstates the case. But, minimally, let's accept this depiction as **Dostoyevsky's challenge**. The Dostoyevskian challenge is a generalisation of van Fraassen style worries about the artificial nature of Bayesian inference mechanism. We should, if there remains any hope for the probabilistic model of belief, be able to mount a reply. Stated somewhat more pedantically, **Dostoyevsky's challenge** amounts to the claim that we can admit beliefs not motivated by our best standards of reasoning. Or perhaps we just have competing standards of reason operating in parallel. Although, even this statement is inappropriate, for surely no standard of reason motivates a contradictory belief. Better to say that a Dostoyevskian (or underground) agent has a plurality of operative standards of belief. As such Dostoyevsky's challenge is a generalisation of van Fraassen's worry, that there is more to belief generation than a mysterious but impulsive probabilistic measure because we have multiple such impulses, and they rarely come amended with numeric values. To show case this fact we consider an experiment under which people profess to beliefs no Bayesian has any business acknowledging.

2.3.2 Ignoring Ignorance

Ignorance, you might think, comes in degrees. For instance, in a state of ignorance we may have a finite range of potential preferences. For example suppose that a bag of 100 marbles is in your possession. You know that 30 marbles are red, and the remaining marbles are either yellow or blue. But since there are 70 remaining, we know that there is an upper bound of 70 blue marbles, if there are no yellow marbles. Alternatively, if we were to expect an equal distribution of blue and yellow marbles, then there would be maximally 35 of each colour. Ignorance is the state in which we remain if we cannot distinguish between the plausibility or probability of these contrary hypotheses.

Halpern reports[26] an experiment involving just such a the situation. In the experiment, if people are prepared to bet on the colour of the ball drawn from the bag, they invariably prefer to bet that the ball withdrawn will be red. There is very little reason to think that this is a well motivated preference, but it implies a belief not accounted for by the Bayesian model of belief formation. Such a tendency to guess under conditions of ignorance can be put down as an attempt to "succeed" in a test environment, or to avoid the appearance of ignorance.

Let's distinguish between the options we have in such a case. We have two unknown parameters B and Y in our setting, these refer to the number of the blue and yellow balls respectively. Allow that $\mathcal{P} = \{\sigma_i^n \mid \sigma_i : B \mapsto [0, .7]\}$, then if we extend our model (Ω, σ) to (Ω, σ_i^m) so that $\sigma_i^m \in \mathcal{P}$ is a probability function which coincides with σ for all valuations, and also specifies a probability assignment for B. On this setting we now have enough information to determine which is the appropriate bet. For, whatever $\sigma_i(B)$, we know that $\sigma_i(Y) = .7$ - $\sigma_i(B)$. In face of the objection that we cannot expect any further information, we should of course concede our ignorance. We have knowledge only of the limit cases which is to say we know that for any agent i there are two probability functions σ^* and σ_* where σ^* ensures that at most $\sigma^*(B) = .7$, and at least $\sigma_*(\phi) = 0$. Call these the upper and lower probabilities respectively. But given that they don't concede ignorance what information to the experiment's participants rely upon? Without an effectively

[25] *Notes From the Underground* pg22

[26] cf. [45]

arbitrary decision rule, or more information we are stuck in a state of ignorance. Observing that people profess specific beliefs and ignore their own ignorance, we should think that they have another motivation for such a confession.

You might think that the Bayesian can account for multiple sources of belief if, instead of simply accepting σ as our belief function, we adopt the view that whenever we ascribe belief to ourselves there is a random choice made to take a particular belief-distribution function from $\mathcal{P}$. However, this does not help elaborate the notion of belief, in fact it shows that on the Bayesian belief is a poorly understood primitive. This last point is made vividly in the basic Bayesian game theory[27] which attempts to cash out the possibility of mixed strategic reasoning by adding a randomising function over a set of probability measures, to represent the beliefs of a more "realistic" agent. This concession says it all. Nothing about the Bayesian setting can account for the inherent plausibility or probability of our particular beliefs. The Bayesian setting is idealised in the sense that it only represents a normative standard of reasoning for once our beliefs have been decided. Hence, as a means of deciding between competing theories, it is effectively useless for our beliefs (having already been decided) will always prejudice us for one theory of another inexplicably.

Allow that the actions participants in our experiment are indicative of a belief. Their decision is evidence of **a bias**. The idea that these decision rules are not arbitrary, is the view that we may have legitimate justifications and explanations which motivate beliefs without directly factoring for our probabilistic data. We call this a **tolerable bias**. We only distinguish between a tolerable bias and a bias, if there is some stake (or inherent value) in the decision which would mandate the unavoidable adoption of some or other bias. So for instance, in the above experiment we might explain our result as a feature of the "classroom" environment of the experiment and the desire of the participants to appear knowledgeable. A tolerable bias is one mandated by the inherent value of the decision.[28] However, it's far from clear that addition of stakes to any particular decision will help us solve such a problem.

2.3.3 The Problem of Mother Mary

We have shown above that decisions under ignorance (in the Bayesian setting) can only be seen as prompted by an ultimately mysterious doxastic impulse inexplicable in the Bayesian setting. If we follow the standard assumptions of decision theory, we might hope to account for such beliefs as being prompted by their effective utility.

L.A.Paul has recently argued[29] that the standard decision theory model cannot account for our reasoning about a variety of life-defining decisions. She cites the example of motherhood, and our inability to calculate the expected value of motherhood. Arguing by analogy with Jackson's famous example of Mary, who is omniscient but colour blind, Paul imagines an expecting mother Mary who knows all there is to know about the objective rates relevant to child rearing e.g. miscarriage, chances of cot-death, learning deficits, life expectancy, etc.... and all the anecdotal reports of ensuing happiness, love, satisfaction and frustration which come with child rearing. Paul argues, that Mary still does not know what it is like to be a mother since no such evidence plausibly prepares you for the event of motherhood i.e no objective probability assignment σ can be seen as a reliable gauge of expected utility. Not for nothing, do most mothers think their child an exception to every rule, or so goes the argument.

In a Bayesian game where the competitors are caught up in some kind "self actualisation" competition, what kind of reasoning would suffice to motivate the claim that having a child is an acceptable, indeed ultimately profitable, strategy? Since no utility assignment can be seen as an

[27] cf. [70]

[28] Typically on the basis of some agreeable generalisation.

[29] [59]

accurate gauge of our ultimate expected return, we should consider rules of decision which are not directly tied to objective probability measures. Call these rationalisations either justifications or explanations. It is by making such actions that we can motivate a decision. If our opponents share our rationalisations, then a justification game can begin. There are two tiers here. On the one hand, we can cite heuristic explanations which purely restrict the probability space, and say nothing of the ultimate payoff - as in the case with heuristic generalisations about, politicians and princes in Machiavellian games. Even assuming that our political opponents are the most fiendish of antagonists, more prone to devious action than the average devil, our choices are still not uniquely determined by a proffered incentive for a particular course of action. In Paul's example, such an action is not sufficient. If we are to motivate a decision we need to cite an explanation which, perhaps restricts the probability space, and necessarily assigns a particular utility function to our considerations such that no other considerations are deemed relevant. This she claims, is not possible.

Reflection

We suggest that an agent's justifications and explanations have to be subjective and agreeable, all the while being effectively independent of our probability measure because if we cite any subjective justifications which rely on our probability data e.g. if we accept a rule of minimal conservative risk, we still need to be able to assign a utility measure over the probability space to determine our course of action, and this is precisely what Paul argues that we cannot do in case of "transformative" events. So in our example you might think that a brute appeal to the "beauty of motherhood" is sufficient to alter the utility measure such that all all lotteries over Ω make the choice to have a child more profitable than the opposite. Of course this motivation only works with a particularly sentimental audience. A consequence of this view is that in order to have any kind of rational decision procedure over whether to enter into any "transformative" experience, we must appeal to an inherently social and local method of decision as societal corroboration is the only real constraint on the plausibility of heuristic decision mechanisms. *We need to pick our bias!* This easily generalises to all cases of reasoning over "transformative firsts". Presuming, as we do, that there are legitimate reasons to have, and refuse to have children, we suggest that this argument indicates a profound failing of traditional decision theory as a model of reasoning.

For many practical cases of reasoning we cite such heuristics justifications primarily as a result of ambiguity aversion. The idea that we can expect the future to conform with typical patterns for the sake of familiarity. It's easier to proceed with life changing decisions if they can be motivated by a norm e.g. taxation inhibits enterprise so no taxes are good.[30] This suggests that our reasoning is, at least, not wholly Bayesian in format. Most decisively, Kahneman and Tversky's famous conjunction fallacy, has shown that people prefer to reason by heuristic rule rather than the probabilistic calculus.[31]. Their experiment presents participants with a suggestive narrative about Linda a philosophy graduate with political leanings, and asks the participants to gauge Linda's likely behaviour. Crucially, participants ignore the axioms of probability in their estimation of Linda's likelihoods prejudiced by their informal expectations.

So far we have elaborated some problems which mutually support a rejection of the purely quantitative approach to belief generation. The trajectory we developed was as follows: we observed with Dostoevsky that there was something too idealised about the Bayesian model of belief, and in particular that this idealisation could not factor well for the actual reasoning of human beings in cases of ignorance and apparent whimsy. We sought to explain the the tendency of human beings to make effectively arbitrary choices, in cases of ignorance as a symptom of imagining a stake in the decision. We then argued that the interpretation of a Bayesian model could not be

[30] For a discussion of this phenomena see [3]

[31] See [1]

rescued by the addition of stakes, since this development involved an ambiguous notion of value. For instance we could disambiguate this notion by suggesting that actions can come in varieties. Some actions can be seen as efforts to restrict the rewards accorded to others or maximise your own reward, and other actions can be seen as efforts to arbitrarily "re-value" the utility function. We saw that neither action would suffice, since neither should be thought of as tracking purely objective evolutions of the utility payouts, since there are actions for which no quantitative data was being factored for, i.e. reasoning under under ignorance or about a "transformative" first.

2.4 Conclusion

In this chapter we have developed the Bayesian model of belief as standardly understood and extended it to encompass a variety of reasoning scenarios. In each case we have pointed to a number of inadequacies of the formal setting. Our criticisms were directed in such a way as to motivate the consideration of qualitative rules of belief formation. In particular we take ourselves to have shown that belief is not well understood as primitive and total function. With this in mind, we continue to the next chapter, wherein we will develop a qualitative account of belief and reasoning based primarily on the justification-logics of Sergei Artemov. Again the goal is to develop a setting in which we can properly couch explanatory reasoning.

Chapter 3

Plausibly True: Qualitative Rankings

Criticism is prejudice made plausible - H.L. Mencken

3.1 Introduction: A Qualitative Model of Belief

In this chapter we shall present a number of logics which purport to be a model for belief formation and reasoning. Importantly, these logics should be thought of as purely qualitative models of belief and inference as we do not directly factor here for any kind of quantitative "degrees of belief". This is not because we have ruled out the importance of probabilistic reasoning, but because we feel that model of reasoning is incomplete without a logical component. This chapter will serve as both a survey of existing formal options and an attempt to model the course of belief generation, and the manner in which a notion of rational belief can be seen as a result of a justificatory process.

The hope is that once we have this notion of reasonable belief we can then apply it to particular abduction problems. We will motivate this form of reasonable belief independently of its application to the grue-hypothesis. We seek a principled answer to the grue-paradox that can be elaborated in a formal setting. In this manner we hope to show that abductive reasoning allows us to track particular kinds of explanatory or justificatory relations. We will use this observation to ultimately sketch a more general theory of explanation.

3.2 Epistemic Logic

We begin with the presentation of a simple Epistemic Logic and slowly augment the syntax and semantics to better capture the subtleties of our reasoning. The reason we begin with epistemic notions is that for all its machinery the Bayesian formalism does not have cogent analysis of knowledge. We can subscribe to the view that knowledge is simply certain belief i.e. that a proposition E is known just when $\sigma(\mathrm{E}) = 1$. But this is evidently false, since no degree of certainty will ensure truth. Imagine an infinite array of points. Throw a dart and you will be probabilistically certain that you have not landed on the 100th point in the array. However, you will not know this to be true. Hence, the Bayesian setting does not provide us with a cogent account of knowledge. Our inability to factor for knowledge in the Bayesian setting is sufficient to disqualify the Bayesian formalism from being a model of our inferential practices. This is both an empirical and normative failing for the Bayesian.

A logic is supposed to be a model of (or often an aid to) reasoning. It is usually a normative model. But assuming we wish to model the practice of scientific inference, which is *de facto* a normative species of inference, we need not be too concerned that logical reasoning is inherently (and perhaps unachievably) normative. A logic of knowledge and belief can at the same time, be empirical if scientists do in fact subscribe to, and act on, the norms of their own discipline. We reason over and about certain claims and counter claims, we argue over the truth or falsity of particular propositions. If for no other reason than brevity, we now provide a minimal language in which to represent this kind of reasoning succinctly.

3.2.1 The Syntax

Allow that $\mathcal{L}_E$ is the language of our logic i.e. the set of all constructible propositions. We let $\mathbf{At} = \{p \mid p \text{ is atomic }\}$, which is just to say **At** is the collection of base propositions from which we can build more complex formulas. An atomic proposition is syntactically simple, a complex proposition is one which is built from atomic propositions by means of the following construction rules. In other words the set $\mathcal{L}_E$ is constructed recursively from operations on atomic sentences.

- If p $\in$**At**, then p $\in \mathcal{L}_E$
- If ϕ and ψ are in $\mathcal{L}_E$, then $\neg\phi$, $\phi \vee \psi$ and Kϕ are also in $\mathcal{L}_E$.

There are no other elements of $\mathcal{L}_E$. We may present $\mathcal{L}_E$ more succinctly as a modal lanaguage which admits formulas of the following shape: We use the following notation to represent the terms occuring within our recursive construction.

$$\phi ::= p \mid \neg\phi \mid \phi \vee \psi \mid \mathrm{K}\phi$$

The usual abbreviations can be defined for convenience, so that the other connectives, i.e. the conjunction ($\wedge$) and the conditional ($\rightarrow$), and the biconditional ($\leftrightarrow$) can be used without incorporating them as primitives. For example; $\phi \wedge \psi := \neg(\neg\phi \vee \neg\psi)$ and $\bot := \phi \wedge \neg\phi$. This is all familiar from classical propositional logic. The only novel feature of $\mathcal{L}_E$ are the formulas of the form Kϕ.

The intuitive reading of these (Kϕ) type of formulas is supposed to be something like *it is known that* ϕ, but such a reading will be clearer in light of the semantics. Before we move on to those details observe that we may define $\overline{K}\phi := \neg\mathrm{K}\neg\phi$, where the former is supposed to be understood as the claim *It is not known that* $\neg\phi$, in other words, ϕ is epistemically possible, or consistent with what is known.

3.2.2 The Interpretation and the Semantics

So we have a language. Now we need to know when it is that we say true things in our language. In particular we need to define truth so as to respect the desired readings of the modal operators K and $\overline{K}$. The idea is that we need to specify the conditions of interpretation for formulas in $\mathcal{L}_E$ and in so doing we follow the ideas of Hintikka[1] by developing a *possible worlds model* of knowledge. The idea is to represent an agent's uncertainty as a range of possible worlds and depict incremental learning as a kind of winnowing of possible options. The more we know about the world the fewer possibilities we entertain. In the immediate, we present only the structural model, leaving further reflections about the cogency of the idea until later.

Definition (A Kripke Structure) Let **At** be as above, then a possible worlds model M = <W, $\sim$, V >, where the following conditions hold:

- W is a non-empty set of consistent possible worlds.
- $\sim$ is a possibly empty relation on the set of possible worlds, i.e. $\sim \subseteq$ W $\times$ W, called an accessibility relation, to denote which worlds are deemed possible from the current world.[2]
- V is a valuation function V: W $\mapsto \wp$(At), so as to denote which atomic propositions are true at each world.

[1] [49]

[2] Traditionally $\sim$ is an equivalence relation for reasons we hope to make clear below.

So the idea is that an agents knowledge is represented by a model. If our agent finds that she knows ϕ, then there is no possible world $\sim$-related to w, in which ϕ is false because to know something to be true, is to know it couldn't be otherwise. This is the central idea behind Hintikka's notion of knowledge. Other kinds of principles suggest themselves. You might think that you are self aware. If you do, you presumably believe that you can reflect on what you know. As such you could insist that if you know ϕ, you know, you know ϕ. These principles can be encoded by the type of relation $\sim$ is, for instance if we wish to ensure that knowledge entails truth we should insist that $\sim$ is a reflexive relation, and that if Kϕ is true at a world w, then ϕ is true at the world w since w $\sim$ w, for all worlds by the reflexivity of the relation.

Such a setting allows to attribute certain kinds of information to any particular agent by elaborating their knowledge-state in terms of possible worlds model with more or less restraints on the relation $\sim$. From here on out, unless we state otherwise, we shall treat $\sim$ as an equivalence relation. We tie these notions together by defining our semantics and elaborating some consequences of this stipulation:

Definition (Truth and Satisfaction) We denote a agent's epistemic situation as a pair (M, w), where M = < W, $\sim$, V >. Then,

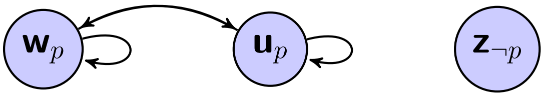
- M, w $\models \phi$ iff $\phi \in$ V(w)
- M, w $\models \neg\phi$ iff M, w $\not\models \phi$
- M, w $\models \phi \vee \psi$ iff M, w $\models \phi$ or M, w $\models \psi$.
- M, w $\models$ Kϕ iff $\forall$w' (w $\sim$ w'), implies M, w' $\models \phi$.

Note that the agent always and only considers possibilities locally, that is to say the relation $\sim$ is always considered with respect to a particular point of consideration i.e. w above. We defined knowledge with respect to the set of all available possibilities, however the dual notion of consistency with our knowledge, i.e. $\overline{K}\phi$ can also be defined in this setting.

$$\overline{K}\phi \text{ iff } \exists w'(w \sim w')\text{and M, w' } \models \phi$$

We often hope that the possibilities we consider are in genuine relation to the real world. If we consider a number of possible worlds, each of which validate identical knowledge claims, we might say these worlds are indistinguishable from our point of view. How can we knowingly distinguish two possible worlds if both validate everything we know to be true? We give an example where $\sim$ is not reflexive.

Example Consider the model M = < W, $\sim$, V > where W = $\{w, u, z\}$, and $\sim = \{(w, w), (w, u), (u, u), (u, w)\}$. Furthermore, assume **At** = $\{p\}$ and p $\notin$ V(z)



Now, consider any agent whose state of information is denoted by this model. Not that p is known, since all worlds $\sim$-related validate p. This is easy to see since no $\sim$-related world validates anything contrary. However consider the world, z and note that M, z $\models \neg$p. Worse recall the semantics for K, and observe that every proposition can be said to be known, since the satisfaction clause is vacuously satisfied. In particular M, z $\models$K$\bot$. This is surely problematic?

To avoid such a situation we minimally stipulate that $\sim$ is a serial relation, then we ensure the existence of some world $\sim$-related to z. Our model ensures that $\sim$ is clearly both symmetric and also vacuously transitive. In fact, it is also reflexive if we restrict our attention to w and u we cannot distinguish the worlds w and u by the propositions they validate - they effectively collapse into one another for all intensive purposes. By definition each world validates the propositional tautologies, this last remark indicates that knowledge is closed under logical consequence. Since every world is taken to be consistent all the theorems of classical logic will be valid at each world, and so known at any world whatsoever, perhaps trivially depending on the structure of the model. This is called the problem of logical omniscience.[3].

3.2.3 The Problem of Logical Omniscience

As you can imagine any reasonable account of human rationality would wish to avoid the conclusion that human agents are logically omniscient. We do not know all the logical consequences of what we do know, we are not unerringly reflective or incessantly deliberative. It is these kind of reasons which prompt Stalnaker to reject this kind of logic of knowledge and belief, and the underlying image of cognition it represents.[4] The centrally offending idea is thought to be the notion that we can characterise knowledge and belief in terms of the conscious attitudes we hold towards a stock of sentences, or propositions, since straightforwardly this characterisation totally elides the notion of implicit or explicitly held beliefs. Worse, it says nothing of the source of our beliefs or knowledge. Both are taken as unexamined primitives, and every agent comes equipped with a stock of beliefs and indeafeasible knowledge.

Consider the semantics of knowledge above. Such a setting makes it quite tempting to think of knowledge as a set of sentences which have been validated by an equivalence class $|w|$ of worlds. As such an agent is ascribed knowledge in so far as we picture them as considering one amongst the set of propositions ϕ where M $|w|$, $\models \phi$. An agent is thought to know a formula just when they are capable of recovering this formula Kϕ from the epistemic position i.e. the model of their knowledge. This, it is supposed, is all there is to knowledge on the sentence storage model. The model lacks the ability to make fine grained distinctions between beliefs which are expressed via extensionally equivalent propositions but are differently evocative. This stems from the operative notion of truth-functional equivalence among propositions. Most importantly it is not obvious what kinds of relations must hold amongst these stock of sentences for us to deem them an appropriate approximation of an individual's beliefs. In short, the notion of knowledge/belief as captured by the "sentence storage" model is too simple.

First observe that our semantics are currently strong enough to prove some simple theorems which characterise knowledge. For instance, Kripke's famous K-axiom follows immediately from the definition of our modal operator. Let w be an arbitrary world in M, such that M, w $\models$ K$(\phi \to \psi) \wedge$ Kϕ, by first conjunct every related world u is such that M, u $\models \phi \to \psi$, and by the second conjunct every world validates ϕ hence by modus ponens, every world u validates ψ, so M, w $\models$ K$\phi \to$ Kψ as desired. This is a simple theorem of any modal logic which ensures that we will come to know the material consequence of our knowledge. This problem becomes very pernicious if we allow that the set of **At** to contain infinite propositions, and insist that each world is a maximally consistent set, as done in developing the canonical model for our logic.[5] Not only do we then, know the material consequences of our knowledge, we know the infinite material consequences of our knowledge!

[3] This semantic feature is encoded in the proof theory by the rule of Necessitation which states that if $\vdash \phi$, then $\vdash$Kϕ

[4] In [87] Stalnaker diagnoses a vulnerability of epistemological/doxastic modesl based on what he calls the *sentence storage* model of belief. This is, he claims, a harmful idealisation.

[5] We name each modal logic by listing the historical names of the axioms it validates. So above we have shown that whatever the restrictions on $\sim$ our semantics validates the K-axiom. As such all modal logics validate the K-axiom and we may state the name of a logic as K?? depending on the other axioms it validates. Below we shall give the historical names of the axioms for clarity.

Seen as an attempt to lend some structure to the "sentence storage" model, Epistemic logic is still too simple. Mandating that our knowledge is closed under logical consequence results in an intolerable idealisation under which agents are depicted as operating with an inappropriate degree of precision. We need further amendments to our model, if we are to hope to better represent the conditions of knowledge and belief.

Knowledge and Structure

In this section we shall (a) canvas in some detail the restrictions we have imposed upon $\sim$ and the consequences had by these restrictions, and (b) elaborate a notion of belief separate from but analogous to the presentation of knowledge.

Technically we have just shown that Knowledge is closed under modus ponens. To fully show that KT45 implies logical omniscience we should also show that knowledge is closed under operations of logical consequence and logical equivalence. The former follows by the Necessitation rule. We quickly show that that (a) if $\phi \leftrightarrow \psi$, then $\mathrm{K}\phi \leftrightarrow \mathrm{K}\psi$. First assume that $\vdash \phi \rightarrow \psi$, then by Necessitation, we get $\mathrm{K}(\phi \rightarrow \psi)$, and by the K-axiom we have $\mathrm{K}\phi \rightarrow \mathrm{K}\psi$. Now it's easy to see that (b) if $\phi \leftrightarrow \psi$, then $\mathrm{K}\phi \leftrightarrow \mathrm{K}\psi$ follows quickly. Since by (a) each conditional in our antecedent proves the appropriate direction for validating our consequent. By adding further structural constraints we might hope to prevent such consequences, or at least give a better (less idealised) characterisation of knowledge. We consider three amendments and observe some emerging theorems.

If we let $\sim$ be a reflexive relation, then every world has a reflexive loop like u, and w in the above example. This has the consequence of validating the following formula: $\mathrm{K}\phi \rightarrow \phi$. This is known as the T Axiom or the Factivity Axiom. Similarly, if we set $\sim$ to be a transitive relation then every world in any model will preserve the following theorem: $\mathrm{K}\phi \rightarrow \mathrm{KK}\phi$. This is known as the 4 Axiom or the Positive Introspection Axiom. Finally, if we insist that $\sim$ is a symmetric relation we will be validating the claim: $\phi \rightarrow \mathrm{K}\overline{K}\phi$. This is known as the B Axiom. These are all elementary results, but the nice feature of these formulas is that they correspond exactly to the condition on $\sim$.[6] That is to say, not only are they entailed by specifications of $\sim$, but if we add them to our axiom set, then they enforce the conditions described.[7] It is for this reason we stipulated the $\sim$ is a equivalence relation.

Now you might question whether these conditions are appropriate for modelling the notion of knowledge. The traditional answer says: They are definitive! A moment of reflection will bring the point home quite clearly; the T-axiom says what is known is true. Hard to dispute, but not impossible. For instance, you might subscribe to a contextual definition of knowledge wherein the conditions of what counts as known, differ with respect to the context of evaluation. Moving from one context to another tightens or respectively loosens such restrictions. In one context I know that France is hexagonal, in another I know that this is strictly speaking false. Arguably this is tolerable definition of knowledge, under which knowledge does not entail truth. The 4-axiom says, that what is known is known to be such. This is less obvious, but true in your better moments. The B-axiom says that for anything which is true, we may come to know it without contradiction. This, to say the least, is very optimistic. Other even less obvious principles can be derived when combining the three conditions.

[6] [72]

[7] Consider Reflexivity - the proof is quick. Assume $\sim$ is reflexive, then pick an arbitrary world w where M, w $\models$ $\mathrm{K}\phi$, then by definition of $\sim$ M, w $\models \phi$. Since w was arbitrary this holds for all worlds. Now the other direction. We prove it by contraposition. Assume $\sim$ is not reflexive. Now there is at least one world w such that $\neg$(w $\sim$ w) let it be that $\phi \notin$ V(w), but is valid in all other worlds related to w. Hence, M, w $\models \mathrm{K}\phi \wedge \neg\phi$. This is a counterexample. Applying contraposition again we get, if $\models \mathrm{K}\phi \rightarrow \phi$, then $\sim$ is reflexive. This completes the proof.

Example A K4B proof. We use a Hilbert-style proof system and assume our modal logic is normal.[8]

n	Formula	Justification
(1)	$\vdash_{K4B}$ p $\rightarrow$ K$\overline{K}$p	(B)
(2)	$\vdash_{K4B}$ $\overline{K}$p $\rightarrow$K$\overline{K}\overline{K}$p	(US)
(3)	$\vdash_{K4B}$ $\overline{K}\overline{K}$p $\rightarrow$ $\overline{K}$p	(4)
(4)	$\vdash_{K4B}$ K($\overline{K}\overline{K}$p $\rightarrow$ $\overline{K}$p)	(Necessitation)
(5)	$\vdash_{K4B}$ K(p $\rightarrow$ q) $\rightarrow$ (Kp $\rightarrow$ Kq)	(K)
(6)	$\vdash_{K4B}$ K$\overline{K}\overline{K}$p $\rightarrow$ K$\overline{K}$p	(US), MP, 4, 5
(7)	$\vdash_{K4B}$ $\overline{K}$p $\rightarrow$ K$\overline{K}$p	(PL), 2, 6
(8)	$\vdash_{K4B}$ $\overline{K}\neg$p $\rightarrow$ K$\overline{K}\neg$p	(US), 7
(9)	$\vdash_{K4B}$ $\neg$K$\overline{K}\neg$p $\rightarrow$ $\neg\overline{K}\neg$p	(PL), 8
(10)	$\vdash_{K4B}$ $\overline{K}\neg\overline{K}\neg$p $\rightarrow$ Kp	(Dual), 9
(11)	$\vdash_{K4B}$ $\overline{K}$K$\neg\neg$p $\rightarrow$ Kp	(Dual), 10
(12)	$\vdash_{K4B}$ $\overline{K}$Kp $\rightarrow$ Kp	(PL), 11
(13)	$\vdash_{K4B}$ $\neg$Kp $\rightarrow$ $\neg\overline{K}$Kp	(PL), 12
(14	$\vdash_{K4B}$ $\neg$Kp $\rightarrow$ K$\neg$Kp	(Dual), 13

We have given an K4B-proof of $\neg$Kp $\rightarrow$ K$\neg$Kp[9] . However, the latter theorem states that for anything we do not know, we know we don't know it. This is presumably false. Consider my belief that I know Edmund Hillary landed on the Moon. To validate the theorem I would have to know, that there is a possibility that my belief was false. However I'm not very diligent, so I can't corroborate the theorem. Nevertheless, it is a theorem of our logic which follows from the assumption that $\sim$ is both symmetric and transitive. Similarly, line (12) above records an axiom which is characteristic of a Euclidean relation $\sim$, and states that if it is consistent to know ϕ, then we know ϕ.[10] Another generous characterisation of an agent. To see the sheer strength of this notion of knowledge consider the following example.

Assume that we can have a not necessarily serial, or reflexive model of a knowledge state where the Moore sentence "Kϕ and $\neg\phi$" is true i.e. M, w $\models$ Kϕ $\wedge$ $\neg\phi$. That is to say, suppose we can claim to know something which is in fact false. You probably do this every day. By the B axiom M, w $\models$ K$\overline{K}\neg\phi$. There are two cases: (1) Suppose there is a world u, such that w $\sim$ u, then by our initial assumption, the latter result and the defition of K, we ensure that M, u $\models$ $\phi \wedge \overline{K}\phi$. By the latter conjunct there is a world z such that M, z $\models$ $\neg\phi$ but by the 4-axiom we know that M, z $\models$ $\phi \wedge \neg\phi$ since $\sim$ is transitive. This is a contradiction. (2) There is no world u $\sim$-related to w. This is technically speaking consistent, but it makes all our knowledge the product of vacuous considerations. In short the B-axiom ensures that we may never learn a false-hood when we try, or we can hold a false-hood, if we never to seek to learn. The latter seems true if undesirable as a norm, but the former is false.[11]

However, all of this is moot if we insist on the reflexivity of $\sim$. We can then derive a contradiction in both cases immediately. For on such an assumption we can never know anything which is false. Alternatively if we relaxed our stipulation we could block the possibility of (2) if we instead insist that our knowledge is consistent i.e. we allow $\neg$K$\perp$ is an axiom. But this assumes a reflectiveness not always present in rational deliberation. In short, this characterisation of knowledge is very idealised, and works best as a model of ideal reasoner at the end of inquiry. At the very least it is important to have these notions defined as a kind of limit case. The broad point here is that we can achieve more or less appropriate characterisations of the notion of knowledge by making

[8] A modal logic is normal if it is closed under the operations of Modus Ponens, Necessitation and Uniform substitution and preserves the Kripke Axiom K and Duality of our operators K and $\overline{K}$ cf. [72]

[9] This is also known as the Negative Introspection Axiom

[10] The 5-axiom

[11] For further discussion of these type of Moore sentences consult [60]

certain structural changes to our model. With this in mind we can continue to define a logic for knowledge and belief.

3.2.4 Epistemic Doxastic Logic

As before we need to specify the details of our language and develop an appropriate semantics. The main difference between notions of knowledge and belief is that there appears to be no sensible definition of belief which validates a version of the T-axiom. We cannot infer from our belief in God, the existence of God. We cannot hope that our belief in a bet is sufficient to ensure our success. These simple observations are decisive.

We now wish to introduce the notion of belief into our model. As before we define a model to capture the notion of belief as a Kripke-relation, we then consider the kinds of principles we wish to hold of the notion of belief. We take as a new primitive Bϕ intended to denote belief in ϕ, and we define a dual notion $\overline{B}\phi := \neg \mathrm{B} \neg \phi$.

Definition (A Naive Epistemic Doxastic Model) M = <W, $\sim$, $\lhd$, V >, where W is the non-empty set of worlds, and $\sim$ is an equivalence relation characterised by the axioms for transitivity, reflexivity and symmetry, and $\lhd$ is our belief relation. Of course V is a valuation function as usual.

One reasonable option is to define $\lhd$ as transitive, but not reflexive, and Serial. Axiomatically we have the following:

1. All the Propositional Axiom
2. All the Epistemic Axioms
3. $\mathrm{B}(\phi \rightarrow \psi) \rightarrow (\mathrm{B}\phi \rightarrow \mathrm{B}\psi)$
4. $\mathrm{B}\phi \rightarrow \mathrm{BB}\phi$
5. $\mathrm{B}\phi \rightarrow \neg\mathrm{B}\neg\phi$.

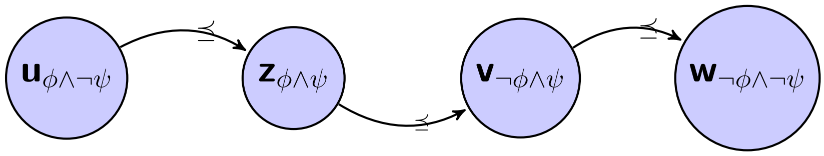
Historically (6) is known as the D-axiom. This D-axiom enforces a serial condition on the relation $\lhd$ of belief. More intuitively the axiom enforces the consistency of our belief, for whenever we believe ϕ, it says, that we cannot believe its converse. This condition is equivalent to insisting that $\neg\mathrm{B}\bot$ is valid at each world. We do not include the Symmetry axiom (i.e. $\phi \rightarrow \mathrm{B}\overline{B}\phi$) because it is perhaps slightly optimistic. It is certainly an ideal of inquiry that we aim at least to consider the possibility of all true things, so adopting the standards of scientific inquiry would at least prompt the belief that we do in fact consider such possibilities. All well and good, so what's the point? How does this setting allow us to represent issues of rational choice? In short, it doesn't.[12] Here belief like knowledge before is an unanalysed primitive. But we need to be able to distinguish between the strength of different beliefs and factor for the source of knowledge and opinion if we have any hope of modelling a process of rational deliberation and choice with respect to underdetermination scenarios.

The Plausibility Relation

Recall that part of the motivation for developing a Bayesian model of belief was to facilitate the representation of uncertainty, particularly uncertain beliefs of varying stripes. You might now wonder how the current setting can capture such uncertainty? In many cases we wish to examine the plausibility of worlds with respect to a particular proposition. Call this the comparison set

[12] It does have the virtue of being a clear model of the type of principles we would like to emerge from a model of rational deliberation.

$||\phi||$, i.e. the set of ϕ-worlds in our model. Consider the following structure.


$\mathbf{u}_{\phi\wedge\neg\psi}$
$\preceq$
$\mathbf{z}_{\phi\wedge\psi}$
$\preceq$
$\mathbf{v}_{\neg\phi\wedge\psi}$
$\preceq$
$\mathbf{w}_{\neg\phi\wedge\neg\psi}$

The thought here is that some beliefs are more plausible than others. A natural question occurs: what are the conditions of a plausibility measure which would be apt to induce belief? We can observe that relative to any particular point of assessment, there is an objective plausibility ordering $\preceq$ with respect to the possible worlds in the model. That is to say (u $\preceq$ z) is to be read as "z is at most as plausible as u" or " u is more, or as equally, plausible as z". With this assumption in mind we can define a set of maximally plausible worlds with respect to w as follows:

MAX $= \{u \in W \mid (u \preceq w')\forall w' \in W\}$

In words, the maximally plausible ϕ-world is the world u such that all worlds w' $\in$ W, u is more plausible than w'. Now we can define the semantics of the belief-operator in terms of the maximally plausible worlds. The overriding assumption is that plausibility is a well founded relation. This is not entirely unmotivated, since if a plausibility ranking is to achieve anything we might wish to be able to distinguish a singular maximally plausible set of worlds. This can then be problematised on the basis that often we are faced with a choice of very plausible for which a decision either way seems arbitrary. In the model above we can see that there are two ϕ-worlds. In principle we could allow that (u $\preceq$ z) and (z $\preceq$ u) i.e that they are both equally plausible, but this is less helpful. Clearly you might want to think that it is more reasonable to at least believe ϕ instead of $\neg\phi$, but is it also reasonable to believe ψ? Insisting that be $\preceq$ be a strict preorder alleviates this choice but also removes any notion of reasonable indecision. This is not obviously a positive feature of the model. What other kind of constraints should be put on $\preceq$ - certainly a plausibility relation ought to be transitive and Lewis[13] suggests it should be reflexive too. More pertinently, how does belief relate to plausibility? Consider the following definition:

Attempted Definition (Primitive Belief)

- M, w $\models$ Bϕ iff $\forall$w' w' $\in$ MAX, then M, w' $\models$ ϕ[14]

This definition assumes MAX is always well defined, and in particular that $\preceq$ is well-founded, hence it is easy to check that $\preceq$ is reflexive and connected. The obvious question is what motivates this plausibility ordering? It ignores all of our available information in that it is defined regardless of our beliefs and knowledge. In other words, we have defined the notion of belief in terms of this primitive plausibility not the other way around. The difficulty is that beliefs are *not primitive*, they are not "so to speak" given, rather they are the result of an inference process. As such, the attempt to define belief in a manner analogous to our definition of knowledge will fail. Not merely because it enforces impossible standards of reasoning, but because the project is fundamentally flawed from the outset. Belief is determined whilst conditional on a theory or wider expectation, any model which fails to capture this subtle point is a poor model of belief. Belief is inherently conditional, and as such there is no inherent plausibility to any particular belief.[15] The plausibility

[13] [63]

[14] Note that it is easy to recover the natural axioms for belief from this definition: For the 4-axiom suppose M, w $\models$ Bϕ. We need to show that $||B\phi|| \in$ MAX, but since $||\phi|| \in$MAX by assumption and $\preceq$ is reflexive $||$B$\phi|| \in$MAX, so M, w $\models$ BBϕ as desired. The other cases are similar.

[15] A notable problem with this view involves the regress prompted by the question: how did we achieve the beliefs our further beliefs are conditional upon? However, we need not insist that beliefs are conditional on other beliefs, belief can be conditional on our knowledge, or dogma. Neither of the latter options need in turn be conditional on anything further if we insist that some knowledge is primitive or basic.

of a belief is conditional on the other available information.[16]

Sellars argues effectively[17] for this point; his strategy is to canvas a wide category of beliefs and undermine the contention that they can be achieved without appeal to inference. For instance he thinks you cannot claim unmediated perceptual belief if you claim perceptual belief of an object's colour, since the ability to speak cogently about colour involves a wider breadth of premises than can be inferred from the immediate perceptual event. Hence, so goes the story, your perceptual belief is conditional on knowledge (or further belief) of the appropriate use of colour vocabulary. The argument is not definitive but highly suggestive. The Sellarsian argument suggests that, we should bar the expression of *primitive* beliefs as incoherent. This is perhaps too strong, but not by much.

3.2.5 Conditional Belief

These last considerations are suggestive but not fully articulated. Worse we have not shown that there is any relation between belief and knowledge. Addressing this second problem is reasonably straightforward and allows us to develop a response to Sellars' concerns. Instead of defining the plausibility relation directly in terms of $\preceq$, we are better off saying that plausibility is determined relative to either what we know and believe. In general we use this section to showcase the sheer expressiveness of our doxastic epistemic logic, but in particular we show how to avoid treating belief *a la* the Bayesians as an unanalysed primitive. First consider how plausibility can be simply defined in terms what we know.

Definition (An Epistemic Plausibility Structure) M = < W, $\sim$ $\preceq$ V > is a model for knowledge in which plausibility is defined with respect to only the class of worlds which fall under the $\sim$-relation. In other words $\preceq$ is a well preorder relation as above, but any $\preceq$-comparable states are also $\sim$-comparable states. In other words we rank only the possibilities which we know to be possible. This avoids the problem we encountered when attempting to define plausibility simply in terms of belief.

Allowing that we might wish to define plausibility with repsect to our working knowledge we should incorporate this in our definition. We define:

$$w \preceq_E w' \text{ iff } w \sim w' \text{ and } w \preceq w'$$

This allows that we define a maximally plausible set of worlds $\text{MAX}_{\preceq_E} = \{u \mid (u \preceq_E w')\forall w' \in W\}$. With this definition in mind we induce a local plausibility structrue from our Epistemic plausibility structure.

Definition (A Local Plausibility Structure) M = < W, $\preceq_E$, V >. Implicit in this model is both the notion of epistemic and plausibility relations. As before $\sim$ is epistemic equivalence relation designed so that an agent is unable to epistemically distinguish between any pair of worlds which validates the facts known to her.

Given such a setting we can articulate a clearer notion of belief, i.e. conditional belief. We elaborate the language $\mathcal{L_{CB}}$ by recursive construction as follows:

$$\phi ::= p \mid \neg\phi \mid \phi \vee \psi \mid B^{\phi}\psi$$

We do not include a K-operator explicitly because we shall show below that this can be derived from the definitions of conditional belief. As standard we do insist on inclusion of $\rightarrow$, $\top$, $\bot$ as

[16] Similar considerations have been developed in [86] wherin he insists' that for any world there is a selection function that determines the maximally plausible world by which our beliefs are determined.

[17] [82]

abbreviations in the usual way. Now we have an option. First we note that the semantics for conditional belief may be straightforwardly defined. Recall that $||\phi||$ is the set of ϕ worlds in our model.

Definition (Conditional Belief)[18] Let $\text{MAX}^{\phi}_{\preceq_E}$ denote the set of maximally plausible ϕ worlds. Then

$M, w \models B^{\phi}\psi$ iff $MAX^{\phi}_{\preceq_E} \subseteq ||\psi||$

The above definition is to be read as saying that ψ is believed plausible conditional on ϕ, or more naturally, if ϕ is believed locally plausible than so too is ψ. Knowledge can be defined as a limit of conditional belief.[19] We can amend this simple definition to respect the fact that the worlds w' feature amongst only the maximally plausible worlds, whether this plausibility relation be defined in terms of $\lhd$ or $\sim$ or neither. Above I suggested we should treat plausibility as defined in terms of $\preceq_E$ as this is the more natural notion. However, we could make plausibility i.e. $\preceq$ a primitive relation such that is defined without regard to either $\sim$ or $\lhd$ and this prompts a slight variation on the semantics. For there is no Kripke-relation by which to define knowledge as a local modality, because $\preceq$ is a global relation.

Definition (A Brute Plausibility Model) M = < S, $\preceq$, $|| \bullet ||$ >. Where S is the set of states (a synonym for worlds) and $\preceq$ is objective plausibility measure over S $\times$ S. While $|| - ||$ is a called a valuation map, which assigns atomic sentences p to a subset of W.

This modification forces a slight amendment to the semantics. On a brute plausibility model we do not have a relation $\sim$ by which to define knowledge, but this can easily be recovered. Hence M, s $\models$ Kϕ iff $\{s \in S \mid ||\phi|| = S\}$ because we only consider the plausibility of worlds which are possible in light of our knowledge. A natural translation suggests itself since we clearly have M, w $\models$ Kϕ $\Leftrightarrow$ $\forall$w'(w $\sim$ w' $\Rightarrow$ w' $\models \phi$) $\Leftrightarrow$ $\{w' \mid M, w \models \phi \text{ and } w \sim w'\} = S \Leftrightarrow M, s \models K\phi$ for any s $\in$ S. Similar remarks serve to cash out a translation for the belief modality B on the assumption that MAX = $\{s \in S \mid$ s $\lhd$ t $\}$. We've already shown above how a Brute Plausibility structure preserves the 4-axiom for belief. It's easy to see that the D-axiom is also validated. We prove this by contraposition. Suppose M, s $\models$ B$\neg\phi$, we need to show that M, s $\models$ $\neg$Bϕ. Suppose towards a contradiction M, s $\models$ Bϕ, then by definition M, s' $\models \phi$ for all states s' in MAX, but by our initial assumption $\neg\phi$ is true in the most plausible worlds in MAX. This is a contradiction, hence M, s $\models$ $\neg$Bϕ as desired.

However, it is less clear that our definition of conditional belief will mandate the appropriate axioms. In fact, as it turns out, the D-axiom fails for conditional belief just when we conditionalise on a proposition ϕ inconsistent with our knowledge. However, this is desirable! We should not entertain propositions which are inconsistent with our knowledge, because this would clearly involve contradicting by the T-axiom. The notion of conditional belief is very powerful. It allows us to express the notions of knowledge and primitive belief as another special case of conditional belief. The idea is natural since it is easy to check:

$$K\psi \text{ iff } B^{\phi}\psi \text{ for all } \phi \in \mathcal{L}_{\mathcal{CB}} \text{ and } B\phi \text{ iff } B^{\top}\phi$$

[18] A Logic of Conditional Belief and Knowledge can be proven sound and complete with respect to Epistemic Plausibility Structures under the following axiom set: The K-axiom for B$^{\phi}\psi$, The T-axiom for K, Kϕ $\rightarrow$B$^{\phi}\psi$, B$^{\phi}\phi$, $\neg$B$^{\phi}\neg\psi$ $\rightarrow$(B$^{\phi\wedge\psi}\theta$ $\leftrightarrow$(B$^{\phi}(\psi \rightarrow \theta)$, B$^{\phi}\psi$ $\rightarrow$KB$^{\phi}\psi$. These results are discussed in [8] and the proof follows the methods described in [72] and [69].

[19] We can also show that w $\models K\phi \leftrightarrow B^{\neg\phi}\bot$. Let $||\phi||_w$ be set of ϕ-worlds $\sim$-related to w. Observe that w $\models$ Kϕ $\Leftrightarrow ||\phi||_w = \{w' \mid w \sim w'\} \Leftrightarrow ||\neg\phi||_w = \emptyset$ by translation $\Leftrightarrow ||\neg\phi|| \subseteq ||\bot|| \Leftrightarrow$ M, w $\models$ B$^{\neg\phi}\bot$.

More pertinently the notion of knowledge as that which survives belief change is very suggestive. What kinds of change? Any change? For instance, we can follow Stalnaker[20] and define a notion of knowledge which survives belief revision with true premises. In addition, the notion of *primitive* belief can be defined as belief conditional on any trivial truth. Furthermore the idea allows us to show that knowledge as a stable concept because it is must survive theory revision with true information. This is called *safe belief* in the literature.[21]

A Derived Definition (Safe Belief)

$M, w \models \Box\psi$ iff $M, w \models B^{\phi}\psi$ for all $\phi \in \mathcal{L}_{\mathcal{CB}}$ where $M, w \models \phi$

This equivalence is in fact a theorem which depends on the definition of safe belief. Considered relationally, we define safe belief in terms a converse plausibility relation. The idea is that ϕ is safely believed just when ϕ is true in all worlds at least as plausible as the world of assessment. In other words, for any world more plausible than the world of assessment, it will turn out that ϕ is true. Hence we may see safe belief as defined in terms of a total preorder $\succeq$-relation, which is to say it is reflexive and connected.

Definition (Safe Belief)

$M, w \models \Box\phi$ iff $M, w' \models \phi$ where $w' \preceq w$

It is now easy to check that the definition of safe belief in terms of conditional belief derives from the reflexivity of $\preceq$. Importantly, the idea of safe belief differs from the traditional characterisation of knowledge is that safe belief is not negatively introspective.[22] We now consider two paradoxes to demonstrate the expressivity and the limitations of our language.

Fitch's Paradox Suppose we are eternally optimistic and believe that it's possible to know any truth, i.e. we take $\phi \rightarrow \Diamond \mathrm{K}\phi$ as an axiom. Now by substitution we have p $\wedge\neg$Kp $\rightarrow \Diamond$K(p $\wedge\neg$Kp).Then since K-distributes over conjunction[23] we get p $\wedge\neg$Kp $\rightarrow \Diamond$(Kp$\wedge$K$\neg$Kp), and so by the T-axiom and the transitivity of entailment we get p $\wedge\neg$Kp $\rightarrow \Diamond$(Kp$\wedge\neg$Kp). The consequent is equivalent to $\Diamond(\perp)$, which ensures by the standard semantics for $\Diamond$ that there is a contradictory world, if there is any true proposition p which we do not know. This is absurd. Hence it would appear that $\phi \rightarrow \mathrm{K}\phi$ is not a validity. The assumption is clearly too generous. [24]

However, if we wish to maintain our assumption, Fitch's paradox offers another lesson. In effect, Fitch's paradox shows for the notion of knowledge what Sellars' arguments show for the notion of belief. Namely, that not every claim is apt to be known, since the procedures by which we achieve knowledge discriminate among reasonable and unreasonable candidate claims. Whereas Sellars argued that not every claim is a candidate for belief, only those claims induced by our background

[20] The above equivalences and the notion of safe knowledge are discussed in [9].

[21] The notion is particularly well described in [77]

[22] Consider the following counterexample M = < W, $\preceq$, V > where $\preceq$ is as usual and we let safe-belief be defined as above. The safe belief relation happens to be both reflexive and transitive. In this particular model we have $\preceq$ = $\{(w,w),(v,v)(v,w)\}$ where W = $\{w,v\}$. Furthermore let $\phi \in$ V(v), hence M, w $\models \neg\phi \wedge \neg\Box\phi$. Clearly M, v $\models \Box\phi$ and so since M, w $\not\models \Box\neg\Box\phi$ this is a counterexample to the Negative Introspection Axiom for safe belief.

[23] To see this we need to show K($\phi\wedge\psi$) $\leftrightarrow$ (K$\phi\wedge$Kψ). For left to right note (1) $\phi\wedge\psi \rightarrow \phi$ is a theorem of propositional logic, so by Necessitation and the K-axiom, we get K($\phi \wedge \psi$)$\rightarrow$Kϕ and similarly for ψ. So by conjunction we (Kϕ $\wedge$ Kψ). For (2) the other direction, we show that since $\phi \rightarrow \psi \rightarrow (\phi \wedge \psi)$ is a propositional validity, we can apply similar reasoning and the K-axiom to ensure that K($\phi \wedge \psi$). Hence K($\phi \wedge \psi$) $\leftrightarrow$ (K$\phi\wedge$Kψ) as desired. Note that we can object to this principle on the basis of a sceptical possibility. Allow that we know that we are in Amsterdam and we're not victims of an evil demon in Hell, but by design we do not know the latter conjunct individually. This is another way to reject Fitch's paradox, but not one we recommend.

[24] This paradox was first discovered by Fitch; relevant background details can be found in Salerno's collection [79]

theory may be fairly described as being believed. We do not get free substitution into the scope of the belief operator any more than we do in the case of knowledge.

Paradox of the Perfect Believer Suppose we want to express the notion that we believe we know ϕ, when in fact we don't know it all. We might try to express this as $\text{BK}\phi \wedge \neg \text{K}\phi$. But then by the negative introspection of K, we get $\text{K}\neg\text{K}\phi$. Now observe that $\text{K}\phi \rightarrow \text{B}^{\psi}\phi$ for any ψ by definition, then $\text{B}^{\psi}\neg\text{K}\phi$ for any ψ including $\top$, so $\text{B}\neg\text{K}\phi$ by definition of primitive belief. Hence, $\text{B}\bot$ in contradiction to the D-axiom. As such we see that nothing we believe known is anything but known. This is absurd. However, the solution is reasonably simple.[25] Switching our notions of knowledge we try to express the thought as follows: $\text{B}\Box\phi \wedge \neg\Box\phi$, and since the safe-belief modality fails to validate the negative introspection axiom. No such contradiction can emerge. Hence the problem arises from a needless conflation of two distinct notions of knowledge.

We might hope that a similar move can alleviate Fitch's paradox but it is less clear how this would work. Douven[26] suggests that we can undermine Fitch's reasoning by blocking the initial substitution on the grounds that we may only substitute consistent claims into the scope of a knowledge operator. He makes the case that this is simply definitive of the nature of knowledge i.e. we only can credibly claim knowledge of a thing, if it is consistent with what we know and believe. These considerations aside the current system or simple extensions of it cannot accommodate the kinds of problems we seek to reason about. In particular we cannot really express notions of explanatory priority. Like the simple Bayesian setting we can only use our current system to stipulate a plausibility relation which arbitrarily describes a greater degree of plausibility to either grue-hypothesis or its converse. It does not provide a means for explaining the adoption of either claim. However, we need not give up hope that there is a good qualitative model of belief and explanatory reasoning in sight. In particular we need to be able to factor for the relation between evidence and a hypothesis, our knowledge and it's source. Douven's solution to the Fitch paradox cements this point, but it is also obvious. It is somewhat too simple to say that belief in the grue-hypothesis is simply conditional on the belief in our collective evidence. This is undoubtedly true, but the claim holds for the green-hypothesis too. In short, we cannot provide particular reasons for our belief and as such we need to adopt a more fine grained account of evidential relevance.

3.3 Justification Logic

We can see this deficit as a symptom of a more general further flaw in our presentation. In particular we see that so far we have only defined knowledge in terms of unavoidably true belief. We have done nothing to discuss the motivation for belief or knowledge. The inability to express such information seems to enforce the conclusion that all claims are candidates for knowledge. However, the Fitch paradox would suggest that we should be able to exclude certain claims as a candidate for knowledge since they cannot be consistently known. Let's begin with the idea that we can say when a particular claim ϕ admits a justification. This approach is owed to work developed on the logic of proof.[27]

Minimally we ought to be able to say if and when a justification is appropriate to ensure the truth of the justified claim in analogy with the concept of proof, wherein we ought to be able to say when a particular truth is provably valid. Similarly, it should be clear to say when a justification, is in turn justified. These and further principles should be expressible, if we are to attempt to

[25] The paradox was first stated by Voorbraak and the idea for the solution is owed to Baltag and Smets in [9]. Indeed a full discussion of relationships between knowledge, conditional belief, safe belief and primitive belief can be found in this paper too.

[26] [25]

[27] For an overview see [6]

think of belief and knowledge as a result of a justification, or particular sequence or amalgamation of justifications.

3.3.1 The Syntax and Axioms

The crucial idea is that we need greater expressive power. We need to be able say when a particular piece of evidence is (a) relevant to the assessment of some or other hypothesis and (b) when said evidence supports the relevant hypothesis. The point is that neither knowledge or belief are primitives, rather they may be seen to arise in response to the receipt of certain kinds of information. To represent the disparity between emergent beliefs or knowledge and their motivation we define a justification logic which allows us to reason about knowledge and the source of our knowledge.

Minimally we need an expressive language. We define $\mathcal{L}_{\mathcal{J}}$ as a recursive construction operating on two tiers of linguistic elements. The first tier, as before, encodes the basic propositional statements, and the second tier encodes particular justification terms.

$$\phi ::= \bot \mid p \mid \neg\phi \mid \phi \vee \psi \mid j : \phi$$
$$j ::= x \mid j \mid \ j + e \mid j \cdot e \mid !j$$

The intuitive idea is to develop a language which can associate justification terms (i.e. reasons) with propositional claims. So we have the following new recursive rules of construction as follows:

- If j, and r are justification terms then j + r, j · r and !j are terms too. Similarly for the justification variables x and y.
- If ϕ and j are a propositional formula and justification term respectively then j:ϕ is a formula.

We let x: ϕ stand for a variable over justification terms for ϕ. An agent without knowledge of any particular justifications should still be able to reason about logical consequence conditional on the existence of some justifications. With these specifications we have the appropriate expressive potential for articulating reasoning about our reasons. The novel feature of this language is that formulas of the form (j: ϕ) are to be read as saying *j is a justification for the belief of ϕ*. The other expressions involving justification terms denote operations on reasons. For instance, the natural operation of combining justifications should obey certain predictable features. For one, if we have a justification for ϕ then any additional justifications can be added without reversing the status of ϕ i.e. it remains justified. The claim is that ϕ should remain justified whatever other reasons we might also consider. We call this the monotonicity condition for obvious reasons. But there are more complicated procedures by which two justification terms are composed. We might wish to utilise justifications in a sequence, so that it turns out that the material consequences of our justified premises also turn out to be justified. Consider the situation where we have a proof of $\phi \rightarrow \psi$, and a proof of ϕ, reason suggests that there exists a proof of ψ. Conversely, consider the claim that all sets are well orderable. We know this follows from the axiom of choice , so we might claim a justification for the conditional, if Choice, then every set is Well-orderable. However we could lack a compelling justification for the axiom of choice, in which case we might never find a justification for Zermelo's well-ordering theorem. This point is made more forcefully if we observe that the well orderable theorem is equivalent to the axiom of choice.[28]

Consider briefly the axioms of *J*, the minimal justification logic axioms. We will discuss each in turn.[29]

A finite specification of the Classical Axioms.

Axiom of Application j:($\phi \rightarrow \psi$)→ (e:ϕ →(j·e):ψ).

[28] cf. [50] for relevant discussion about the status of how to justify the axiom of choice.

[29] Our discussion draws from the following papers: [4], [**?**], and [29]

Axiom of Monotonicity (j:$\phi \rightarrow$ (j+e): ϕ).

Consider the Application Axiom, this is clearly related to the K-axiom of the standard modal logics. Importantly, the axiom ensures that we can justify the material consequences of our justified claims. This, for many reasons, seems like a minimal standard for reasoning. It is easily prompted by considerations regarding the burden of proof and the expectation of consistency. While intuitively it is much less objectionable than the K-axiom, you can foresee problems if we allow that certain species of justification terms are directly incompatible. Perhaps fortunately this possibility is not even expressible in our setting.

On the more positive side we may add other axioms to facilitate typical features of justification and the process of justification. So for instance, we might wish to insist that a justified proposition implies the truth of that proposition, or that a justified proposition can in turn be justified. This former idea recalls both the T-axiom of epistemic logic and the idea that a proof is a sufficient warrant for accepting the truth of the proven claim. The latter claim recalls the positive introspection axiom of epistemic logic, and the more intuitive idea that if there is a proof j of ϕ there is a proof, that j is a proof of ϕ. We express both conditions as follows:

Factivity-Axiom j:$\phi \rightarrow \phi$

Exponentiation j:$\phi \rightarrow$!j:j:ϕ

With the addition of these axioms we can show that due to the Internalisation rule, any result proven in our justification logic can be expressly said to be proven. In short the justification logic encodes its own proofs as justification terms. In the proof theory we insist that our logic J is closed under the operations of modus ponens, and similarly we should insist that for any valid theorem, there is a justification term corresponding to that theorem i.e. if $\vdash_J \phi$ then $\exists$j, such that j:ϕ. This is clearly true if every proof is to count as a justification term. We will show this below. But first note that we assume the follow rule.

Axiom Necessitation $\vdash_J j : \phi$ if ϕ is an axiom and j is a justification term

Theorem (The Lifting Lemma) If $\overrightarrow{j}$: Φ, $\Delta \vdash_J \psi$ then there is a proof (justification) polynomial t($\overrightarrow{x}$ $\overrightarrow{y}$) such that

$$\overrightarrow{j}:\ \Phi,\ \overrightarrow{y}{:}\Delta \vdash_J \mathrm{t}(\overrightarrow{j},\ \overrightarrow{y})\ {:}\psi.$$[30]

Proof: The proof is by induction on the structure of the derivation $\overrightarrow{j}$: Φ, $\Delta \vdash_J \psi$. Case (1) Let $\psi = \lambda$ be derived as a result of modus ponens from $\tau \rightarrow \lambda$, then by our induction hypothesis there are justification terms u($\overrightarrow{j}$, $\overrightarrow{y}$) and v($\overrightarrow{j}$, $\overrightarrow{y}$) such that u($\overrightarrow{j}$, $\overrightarrow{y}$): $(\tau \rightarrow \lambda)$ and v($\overrightarrow{j}$, $\overrightarrow{y}$) : τ and both are derivable from our antecedent $\overrightarrow{j}$: Φ, $\overrightarrow{y}$:Δ. Then to derive our result we need only apply the **Application** axiom, to derive $\overrightarrow{j}$: Φ, $\overrightarrow{y}$:$\Delta \vdash_J$ (u $\cdot$ v) ψ. So to complete the case we simply define t := (u $\cdot$ v). Case (2) If ψ = j:$\tau \in \overrightarrow{j}$: Φ, then pick t := !j and apply **Exponentiation** on our antecedent. Case(3) If $\psi = \delta_i \in \Delta$, then we pick t := y_i where y_i is among the stock justification variables. If ψ is an axiom, then pick a fresh justification constant c, such that an application of the **Axiom Necessitation** rule ensures c : Ax. Case (4) If ψ is derived by **Axiom Necessitation**, then ψ = c : Ax, for some fresh constant c and an axiom. Use **Exponentiation** to derive the result c: Ax $\rightarrow$!c:c: Ax, , and then apply modus ponens to get !c : Ax and finally fix t := !c $\dashv$

[30] The vector notation refers to the set of possible justification terms indexed to particular propositions in, for instance, Φ.

As a corollary we get the following **Internalisation** rule which states: if $\vdash_J \psi$, then $\vdash_J$ j : ψ for some justification term j which we can always find. This should be evocative of the Necessitation rule of standard modal logics. More generally, the relation between Justification logics and Epistemic logics should be clear. We've effectively swapped primitives. Instead of treating knowledge as our primitive we take the notion of justification to be understood thereby allowing us to simulate the conditions of knowledge in terms of the results a justification prompts. By similar reasoning we might hope to capture the notion of belief as a special case of weak(er), perhaps non-factive, justifications.[31] As before, if we wish to capture any of these nice axioms in a semantic setting we need to define the notion of truth in a model.

3.3.2 The Model and the Semantics

We can define a model for justification logic in analogy with the development of a Kripke structure. The only additional requirement is that we specify an extra constraint so that particular justification terms can feature legitimately in our reasoning at particular worlds. In short, we have to specify at which worlds particular justifications are relevant. To do this we add an evidence function $\mathcal{E}$ which maps states and justifications to sets of propositional formulas.

Definition (An Justification Model) $M_{CS} = <$ W, $\sqsubseteq$, $\mathcal{E}$, V $>$ where W is a set of worlds and $\sqsubseteq$ is a reflexive relation. Strictly speaking $\mathcal{E}$ is a function from pairs of justification terms and propositions into the powerset of W. We often say $\mathcal{E}$(j, ϕ) $\subseteq$ W for all available justification terms and propositions ϕ as restricted by a CS-function.[32]. Of course V is a valuation function as usual.

The semantics for the boolean connectives is as usual we need only specify the semantics for our modal operator.

Definition (Truth and Satisfaction)
M, w $\models$ j:ϕ iff (1) $\forall$w' (w $\sqsubseteq$ w') M, w' $\models \phi$ and (2) w $\in \mathcal{E}(j, \phi)$.

Of course, this semantics happily validates the the factivity axiom of justification logic by the reflexivity of $\sqsubseteq$. But we need to add further constraints to $\mathcal{E}$ if we are to recover the axioms of application and monotonicity. In particular for monotonicity to be preserved we need to specify that the relevance of some justification to a particular propositional claim is inherited by any combination of reasons which includes the initial justification. Similarly, to preserve the justification of implied truths we need to insist that evidentiary relevance is inherited across material implication.

Justification of Implication $\mathcal{E}(j, \phi \rightarrow \psi) \cap \mathcal{E}(r, \phi) \subseteq \mathcal{E}(j \cdot r, \psi)$

Monotonicity of Relevance $\mathcal{E}(j, \phi) \cup \mathcal{E}(r, \phi) \subseteq \mathcal{E}(j + r, \phi)$

These and various other closure conditions on $\mathcal{E}$ can be imposed so as to capture the intuitive axioms. The virtue of this setting is that it allows us to capture a notion of knowledge previously inexpressible. But in particular, this setting displays the nature in which we might hope to address the grue-problem. Which is to say that the grue-hypothesis presents us with an abduction problem, wherein we are forced to choose between two options when neither is obviously justified by our available information. The grue-reasoning proceeds by appeal to a particular kind of evidence function $\mathcal{E}^*$. The challenge is to say if, and why, the $\mathcal{E}^*$-worlds are less plausible than the other worlds $\mathcal{E}$-worlds. That is to say an abduction problem of the grue-type is essentially a question regarding the plausibility of evidentiary links. As such the real problem amounts to the discovery of when particular information is relevant to the assessment of a particular proposition

[31] We will return to this idea below.

[32] This is a meta-linguistic function which assigns the justification terms to particular propositions. The idea is that there are only so many grammatically cogent pairs (j;ϕ) and the evidence function $\mathcal{E}$ must minimally respect the constraints of grammar. We shall discuss this idea further below.

Admissible Justifications

However, before we model the situation, we have to decide which justification claims can be modelled given the constraints imposed by the CS-function. That is to say, not all justification claims will be grammatically cogent. You might think that the claim that all emeralds are green, cannot support the further claim that all emeralds are grue i.e. that the notion of dual colour instantiation is incoherent, or at least incomprehensible given the strictures of contextually appropriate utterance. To meet this restraint Artemov[33] defines a meta-linguistic function called the constant-specification. The idea is that every agent has limited access to justificatory information, so we define a partial function CS: $\mathcal{L}_{\mathcal{J}} \upharpoonright \{j \mid$ j is a justification$\} \mapsto \mathcal{L}_{\mathcal{J}} \upharpoonright \{\phi \mid \phi$ is a proposition $\}$. We can think of this function as a resource distribution for an agent which underwrites the cogency of the claim that j: ϕ for each proposition he takes to be justified. It is a way in which to add contextual parameters to each model. As seen above Artemov does not make the CS-function explicit in the language, but chooses instead to add the information directly to the model by of an Evidence assignment.[34]

Definition (The full CS function) The full CS-function determines a set of formulas j_1 $:\mathrm{Ax}_1 .... \mathrm{j}_n : \mathrm{Ax}_n$ for all instances of each axiom in our logic. CS is injective if there is at most one justification for each instance of every axiom. Every proof in J naturally generates the CS function corresponding to a justification for each step in the proof by the **Internalisation** rule.

To validate basic logical reasoning we need to specify CS in such a way that for every instance of the axiom-schemas in J we have an appropriate distribution of justification term to validate our logical reasoning.[35] Interestingly the model encodes this information as factual information, that is to say a justification claim is true at world, just when j is actually a justification for ϕ. This is a hard-line view of the justification relation as one which arguably tracks a relation of dependence between the justification and the claim. For instance, the justification of our logical axioms is prompted because we may observe the relation of entailment between our premises and our conclusions. This feature of justification is worth bearing in mind if we recall the relation between our best information and the grue-hypothesis. Supposing that there are such fine-grained relations of dependence, the discovery of such a relation would allow us to confirm or deny the grue-hypothesis. The model above hard-codes the existence of these dependence relations on, I think, the very reasonable assumption that they do in fact exist.[36]

So far in the justification logic setting we have only described a model appropriately assigned to be the mental state of some particular agent. As such all the justificatory reasoning is, so to speak, of a subjective nature. We have had the opportunity to generalise the setting to multiple agents, and therefore supply widely agreeable justification relations. An open question is whether these relations hard code objective or subjective species of justification. If the former, then we are faced with the burden of defending the institution of these relationships amongst our beliefs and expectations. If the latter then we face a smaller burden in so far as we need only explain our adoption of certain inference rules, or novel information. In the multi-agent case we might insist on the defining the semantics of j:ϕ by quantifying over all the evidence functions. Enforcing the view that a justification exists in a multi-agent setting just when it is held by all the agents therein.

[33] [5]

[34] In [7] the authors encode this information explicitly in the language by defining a primitive expression j $\ggg \phi$ which works to much the same effect, but is defined at both the syntactic-meta level and the semantic level

[35] A logic is called **axiomatically appropriate** just when each instance of its axiom schemas are justified.

[36] We defer further considerations about the discovery of such relations until our next chapter.

3.3.3 Logical Omniscience: A Solution

In this setting we can naturally address the logical omniscience problem. Since $\mathcal{E}$ is defined in terms of the CS-function, and the closure principles of $\mathcal{E}$ ensure the justification of the consequences of our justified premises, we can halt the possibility that every theorem is justified. The method is simply to define the CS-function such that particular claims cannot be said to be cogently justified. We may think of this operation as an inherent bias and unwillingness to follow through with logical thinking. So for instance, if all the facts, point to the non-existence of a benevolent diety, we might still find a believer who accepts all the facts but doesn't subscribe to atheism, due to his utter incomprehension of the conclusion. This is articulated in our setting by supplying the agent with a limited evidence function so that it never includes a justification for the atheist thesis.

This problem is also raised (albeit in a different guise) by Kripke in his discussion of substitution in belief contexts.[37]. The argument is presented as standard. We may hold beliefs about Cicero's attributes and deny the belief that Tully holds the same attributes. This is presented as a puzzle for any model of belief, because (presumably) we do not wish to deny validity of Liebniz' law. This problem is now easily resolved. On the one hand we have a justification for our beliefs about Cicero, but simply lack a justification for any such beliefs about Tully. This can be thought of as a limitation of the CS function such that we cannot find a justification j for the claim (j): Cicero = Tully. The core idea here is that we can understand beliefs to be derivative of our justifications. Similar ideas apply to the difficulties of knowledge. Consider how this greater expressiveness also allows us to motivate Igor Douven's solution to the Fitch paradox if we restrict the CS function such that no Moorean sentences are ever justified.

3.3.4 Belief via Justification

Recall our logic KT4, we argued above that such a logic could not be said to mimic the appropriate properties of belief, since KT4 $\vdash$ B$\phi \rightarrow \phi$. Our job now is to show how to recover a notion of *justified belief* by deriving model of KT4 from a justification model of JT4. First we note that there is a soundness and completeness result for the logic KT4. The proof is pretty standard [38], so we reproduce the proof in the appendix only to show the connection between justification logics and Doxastic or Epistemic logic. What follows will hold for both Doxastic and Epistemic frames. Furthermore we shall show that there is a finite model of any consistent set of KT4 sentences. But the important result for our current purposes is that we can take any set of justification logic sentences and recover a set of doxastic claims. These can be intuitively thought of as justified beliefs, since they have been recovered from a justification logic setting.

Recovering Belief from Justifications

Having shown that there is a completeness theorem for the standard modal logics of belief, and analogously knowledge. It remains to show that each valid theorem of our doxastic logic can also be understood as a result of the appropriate justification logic. The thought here is that each sentence of our doxastic logic has a hidden structure which has not been revealed. To show this we define a translation function t: $\mathcal{L}_{\mathcal{J}\text{T4}} \mapsto \mathcal{L}_{\text{KT4}}$

$$\begin{gathered} t(p) = p \\ t(\neg\phi) = \neg(t(\phi)) \\ t(\phi \vee \psi) = t(\phi) \vee t(\psi) \\ t(j : \phi) = B(t(\phi)) \end{gathered}$$

[37] In [56]

[38] See for example [72]

This mapping is called the *forgetful functor*.[39] Melvin Fitting[40] discusses its usage with respect to the notions of explicit and implicit knowledge. Where justification logic can be seen encoding a very explicit notion of knowledge, and direct epistemic logic suffices only to articulate claims for which we have an unexamined reason to believe true.

Proposition (Preservation) The forgetful functor maps theorems of JT to theorems of KT. Likewise it maps theorems of JT4 to theorems of KT4 and theorems of KT45 to theorems of JT45.[41]

The result is straightforward but it highlights a neat connection between our beliefs and our justifications. Furthermore, it allows for a tolerable notion of truth-entailing beliefs. This goes a long way to rehabilitating some of mentioned failures of models which only allow for the expression of *primitive* beliefs.

Recovering Justifications from Beliefs

You might now wonder whether given a set of beliefs we can discover a justification for those beliefs. Indeed we can. We shall prove this result in the appendix, but the crucial notions are as follows:

Definition (Realisation) If we let ϕ be a formula in our doxastic logic. A realisation of ϕ is a formula in the appropriate justification logic, that results from applying a realisation function to every subformula B$\psi \in \phi$. A realisation is *normal* if negative occurrences of B are replaced with distinct justification variables (which are always part of the language of our justification logic regardless).

In short, a CS function can be found to ensure the success of each realisation. The technicalities of this definition only become important when proving the next theorem.

Theorem (The Realisation Theorem) If ϕ is a theorem of one of KT, KT4 or KT45, there is some normal realisation of ϕ that is a theorem of JT, JT4, or JT45 respectively.

This latter result was proven syntactically by Artemov and semantically by Fitting.[42] By way of example, we say that (j:ϕ $\rightarrow$!j:j:ϕ) = r(Bϕ $\rightarrow$BBϕ). Where r is a realisation function that takes as input validity in the doxastic epistemic logic, and returns a validity in our justification logic. Another example of a KT4-theorem where (x:$\phi$$\vee$y:$\psi$)$\rightarrow$ (j$\cdot$!x + e$\cdot$!y):(x:$\phi$$\vee$y:$\psi$)) = r(B$\phi$$\vee$B$\psi$) $\rightarrow$ B(B$\phi$$\vee$B$\psi$)). These examples were taken from Artemov, the idea in the latter realisation is that the constants *j* and *e* are from the CS-function so that they stand as justification for the major premises, while for the classical axioms x:ϕ $\rightarrow$(x:$\phi$$\vee$y:$\psi$). So while the theorem receives a more complicated validation in the justification logic setting, we can nevertheless recover a valid theorem of our doxastic logic by means of a (yet to be) specified method of translation. The realisation theorem states that we can work a normalised "translation" in the opposite direction of the forgetful

[39] Bryan Renne uses this translation to prove the consistency of the basic theories of justification logic. [17]pg122 Theorem 3.14. This goes towards proving a completeness result for each justification logic via a canonical model construction. See his Theorem 3.22 for details. Alternatively see Aretemov's overview paper [**?**]

[40] In [30]

[41] *Proof* The proof proceeds in two steps (i) we fix a formula ϕ valid in (for instance) JT4, and proceed to show that for any subformulas of ϕ the translation function works, ultimately applying the translation function to ϕ itself i.e. t(ϕ) $\leftrightarrow$ t(ψ) for all $\psi \in \phi$. Now we show that for any valid formula ϕ in JT4, we can find a model of KT4 which validates t(ϕ). The proof is by contraposition, suppose that M$_K$, w $\nvDash_{KT4}$ t(ϕ) for some w $\in$ W, we need to show that M$_J$ similarly invalidates ϕ. So we define an evidence function $\mathcal{E}$ such that by adding $\mathcal{E}$ to M$_K$ creates a model M$_J$ = < W, R, $\mathcal{E}$ V > where we preserve the truths at every world v $\in$ W. It is now a simple matter to prove that M$_K$, v $\models_{KT4}$ t(τ) iff M$_J$, v $\models$ τ. The proof is by induction as expected and it ensures (given our assumption) that M$_J$ w $\nvDash_{JT4}$ ϕ as desired $\dashv$

[42] In [4] and [29], respectively. There is also nice operator-elimination theorem in [30].

functor. In a sense we discover the greater structure underlying each belief. [43] We have given an example of one normal realisation to give a flavour of the idea, but the more significant thought is that all basic modal logics can now be thought of as short hand for a much more involved process of justification and emergent justified belief.

Justification Relations: Perfectly Relevant?

To summarise the issues so far; we have examined two formal settings which attempt to cash out the notions of belief or knowledge. Both the quantitative Bayesian, and the qualitative Epistemic Doxastic logic settings failed to adequately represent the notions of belief and knowledge in so far as both treated belief as a unexamined primitive. Slight nuance was added by the inclusion of somewhat natural axioms but ultimately both approaches suffered, since they made no mention of the source of our beliefs or the manner in which we come to believe certain claims over others.

This deficit can be addressed in a purely formal setting by adding (as in justification logic) an evidence function $\mathcal{E}$, but the really interesting notions in epistemology involves explaining how such an evidence function arises. These evidence relations encode the relevance of our premises to certain conclusions, so it is vital to suggest a manner in which we can discover how and why certain conclusions are prompted by particular kinds of evidence, and not others. In the next chapter we attempt to provide a theory as to (a) how such connections can be discovered and (b) why such connections perform a vital role in epistemology.

3.4 Conclusion

The models of belief and knowledge elaborated in this chapter point to a common problem. We have not distinguished adequately between the source of our beliefs and knowledge. We have argued that the notion of belief (knowledge) and its attainment is better understood if we explicitly factor for the source our (belief) knowledge. The importance of our current suggestion relates more to the notion that the structure of our information state is inadequately represented if it leaves out a mechanism for distinguishing between the justification of our information, and (perhaps more importantly) a priority ordering on the species of justification deployed. These are the kind of considerations that underlie any solution to an underdetermination problem.

[43] We will prove this result in an appendix.

Chapter 4

True: Structural Commitments

> *Quine taught us some time ago (though not with these words), the metaphysician's task of describing the structure of logical space - the space of all possible worlds - is not so easily separated from the scientist's task of locating the actual world in it.*[1] - Robert Stalnaker

4.1 Introduction: Species of Evidence

In this chapter we shall attempt to unite the foregoing considerations and address the problem of underdetermination which afflicts traditional models of rational belief. First we shall set the scene and demonstrate the pervasive effect of underdetermination problems by discussing abduction and the construction of Gettier cases. We will then elaborate informally a means to develop solutions to abductive problems. The core idea is to distinguish between justificatory claims and the source of our justification. The observation that there are multiple kinds of justification relation defined to be *dependent* on different sources of evidence prompts the idea that no formal epistemology of justification should be elaborated with an attendant formal metaphysics of dependence. The rest of the chapter will showcase an attempt to apply this method to Benacceraf's famous multiple reduction argument in the philosophy of mathematics.

4.2 Abduction

Suppose you are kidnapped and the culprits explain that they seek a ransom, but you secretly believe it to be a prank in prelude to a surprise party. This is a case of abduction.

Intially formulated by C.S. Peirce[2] as a form of explanatory inference the notion of abductive reasoning was supposed to fill the gap between inductive and deductive reasoning. Schematised as follows:

1. You observe the surprising fact ψ.
2. If ϕ were true, then ψ would also be true, as a matter of course.
3. Hence, infer that ϕ is true.

It would be fallacious to think that the truth of ϕ follows deductively from the observation of ψ, but nevertheless this pattern of reasoning is often plausible. Consider our kidnapping case. We have the surprising fact that you have been snatched away from home sweet home. Both the fact that you were worth a ransom and that your friends tend to be pranksters, would explain your current circumstances. But how can we choose which is the better explanation? If we can determine an answer then we should be able to accept one our options as correct. In this way any

[1] [88]pg234

[2] cf. [26]

underdetermination problem may be presented as an abductive problem.

The ideas which shall entertain us in the following section are primarily indebted to Boutilier and Becher, and Soler-Toscano and Velazquez-Queseda[3] We shall aim to keep this discussion relatively informal so as not to commit to any particular formal representation of the issues. In this approach we follow Soler-Toscano and Velazquez-Quesada. We shall argue that abductive problems such as the one described above, are prevalent in our everyday reasoning and such problems create difficulties which undermine the formal notions of belief and knowledge.

We then attempt to show systematically a procedure for solving abductive problems and thereby mitigating the damage to our epistemic notions. Our procedure involves the examination of the role of structural information about the world[4] in our reasoning. In particular how we may use certain types of structural connections to motivate particular solutions to an abductive problem. But first we consider the Gettier problem, this will be used to motivate the discussion of abduction.

4.2.1 The Gettier Problem

Abduction problems allow us to problematise the notions of knowledge and belief. Consider Bob who believes himself to be in possession of twenty euros. Observe that this is true, and that he is justified in his belief because he recalls placing it in his wallet earlier that day. However, unbeknownst to Bob the dastardly Moriarty has replaced his wallet with a convincing fake of comparable content including twenty euros. Gettier argues that despite the fact that Bob's belief was justified and true, he cannot claim to know he possesses twenty euro.

We note two things about this scenario: (1) If we accept Gettier's argument then when faced with the choice between the *My-wallet* and *Moriarty's wallet* we are insisting that we may only achieve knowledge just when the *Moriarity's wallet* is impossible i.e. that we cannot later learn that Moriarty tricked us. In other words for anything to be known our belief must follow by necessity from the justification we develop on the basis of our premises. Seen alongside the grue-problem we can argue that the grue-case is only paradoxical because we conflate the notion of confirmation with the notion of knowledge. Both the grue-hypothesis and the emerald-hypothesis are equally confirmed by our evidence. This is not in itself problematic, but just seems so if we expect our methods of confirmation to converge (without exception) on one particular conclusion. (2) From the minimal premise that knowledge is the desired goal of reasoning we may draw the conclusion that justifications should provide a sufficient conditions for the acceptance of the justified claim. The need to exclude the *Moriarty case* would suggest that a justification need also be a necessary condition. Hence the Gettier argument seems to motivate the requirement that justification ought to provide necessary and sufficient reason for the acceptance our conclusion. This is not a solution, since the move is illegitimate by the fact that it makes JTB-definitive of knowledge and this begs the question on the table. Namely, is knowledge justified true belief?

We can also represent a Gettier problem which undermines even the notion that safe-belief is prompted by justified true belief. We take the example from the paper of Baltag *et al*[5] in which they develop a model to include operators for knowledge, belief and justification. We treat this according to their standard model: M = $< W, \sim, \preceq, E, V >$ where W is the set possible worlds, V is a valuation map, $\sim$ is an equivalence relation for knowledge and $\geq$ is a plausibility preorder as usual. We adopt the notation that $[[\phi]]$ is the set of ϕ-worlds determined by V in W. Belief is defined in terms of the plausibility order $\leq$, while E is their evidence function. In addition to the usual modal operators we have justification terms and an admissibility predicate relating

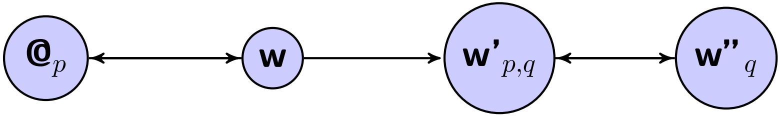
[3] In [16] and [84]

[4] e.g. its causal, mereological or supervenience structure

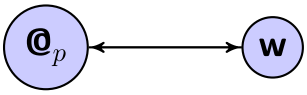
[5] [9]

justification terms with propositional formulas.

Imagine an agent Alice hears of our plight because Moriarty boasts of his evil plot telling her he stole our wallet and replaced it. Unbeknownst to him he has mistakenly replaced our wallet with his own which contains 20euros, before he threw our wallet in the rubbish bin. Let p be defined as "I have 20 euro" and q similarly as the claim"Moriarity has 20 euro". Moriarty boasts of his heist and his comparable wealth of 20euro. Being a reasonable person Alice accepts his testimonial evidence, which implies (a) she finds his evidence admissible and (b) accepts it. Let the actual world be denoted @, then we represent the facts as follows. We omit the arrows ensured by the reflexivity and transitivity.



The plausibility of the worlds goes from right to left with w' and w" being the most plausible worlds in our model. Hence immediately we can see that Alice at the actual world believes either myself or Moriarty have 20e. So M, @ $\models B_a(p \vee q) \wedge (p \vee q)$. Which is to say that Alice believes a true fact since I do in fact possess 20euro as a consequence of Moriarity's foolish plan. Furthermore allow that the Evidence function E is such that testimonial evidence is always considered an apt justification for the claim testified, i.e. we have [j : q], then since all the axioms of classical reasoning are also justified Alice holds that [x : q $\rightarrow$ (p $\vee$ q)] where x denotes the justification variable for the classical right weakening axiom. But then of course by the application axiom there is justification for our conclusion namely that [x·j :(p$\vee$ q)]. In short the initial justification stems from Moriarty's testimony, and the latter two follow as a rule of logic, or by an operation of evidence combination. As such she can derive a justification for her true belief B(p $\vee$ q), but in no sense does Alice know a true fact because M, @ $\models \neg K(p \vee q)$. Worse still she does not even have safe belief in the claim that Moriarty has any money, since an update with the true information that Moriarty has no money, would ensure that all the q-worlds are removed.



Resulting in a picture where the most plausible worlds serve to invalidate Alice's testimonial justification for her belief. Worse she neither believes q or p and M, @ $\not\models \Box_a(p \vee q)$ since despite the fact that @ validates the disjunction, w is equally plausible and invalidates the disjunction contrary to the hypothesis that we have safe belief. In any case, this is a counterexample to the definition of knowledge in terms of justified true belief.[6]

The insight that this suggests is that knowledge as defined by the K-operator is a far more idealised concept than we had previously thought. In practice we may never attain absolute knowledge of the kind required to defeat every Gettier case, but that doesn't do anything to take away the confidence accrued to reliable reasoning. The Gettier problems suggest that we should not try to account for the reliability of reasoning by simply taking our best estimates as infallible, or true because "known". But we do need some kind of measure by which to prioritise certain lines of reasoning. Otherwise any and all intellectual activity is inexplicable, given the judicious paralysis we would face in any abduction problem.

[6] If this example seems too artificial, consider the real life Gettier case involving the communication between Galileo and Kepler which has passed into the folklore of science. Upon discovering Saturn's rings Galileo sends Kepler notice of this discovery in terms of an anagram, thereby patenting the discovery. Kepler (incorrectly) deciphers this anagram to mean that Mars has two moons. The later claim is also true, but would only be discovered two hundred years later. An odd coincidence. Even stranger when Galileo sends news of his next discovery that Venus has phases i.e. that it must be going around the sun. Kepler again deciphers this claim incorrectly to mean that Jupiter has a red spot. This is again true, but not yet discovered. Note the presence of true, justified belief and ask yourself whether Kepler really knew either of his conclusions?

Underdetermination problems are horrendously toxic. Every proposition you care to mention has a gerrymandered competitor by Quinean reasoning. Every single underdetermination problem can be transformed into a Gettier problem! Observe that we are justified in thinking that the emerald we buy tomorrow will be green - we are justified in this belief by the law that all emeralds are green, and better yet, it is true that the emerald will be green. However Gettier's case stems from the fact that we can't be said to know that that the emerald we see tomorrow will be green. Since we are unaware that tomorrow's emerald is rendered green by the time-indexed property of grue-gems. Without providing a reason to reject underdetermination arguments, we risk destabilising the notions of knowledge and belief altogether.

4.2.2 An Abduction Problem/Solution

Let T be our theory[7] , and $\vdash$, be our consequence relation. We have an abductive problem when there is a contention χ before us which is not entailed by our theory. Two potential versions suggest themselves. In one case we also know that T $\vdash \neg\chi$, and in the other T $\nvdash \neg\chi$. In the latter case we can imagine the grue-paradox as a good example. Intuitively, we wish to able to conclude that the emerald-hypothesis is true, but we have no way of proving our contention either way. In the former case we might imagine the case of an informal paradox, wherein we have arrived at a unintuitive conclusion, i.e. T $\vdash \chi$, but we seek to motivate T $\vdash \neg\chi$ as this is the more plausible result. These kind of cases crop up in philosophy all the time. Assume for the sake of argument a premise ϕ such that T$\cup\{\phi\} \vdash \chi$. This is not a *reductio* argument against ϕ since no contradiction need emerge, nevertheless you feel obliged to adapt your theory in such a manner so that $\neg\chi$ comes out as the more plausible result. That is to say, you disbelieve χ despite evidence, and seek to amend the theory so as to promote a belief in $\neg\chi$. A number of amendments can achieve this result. Either you add further information to your theory such that $\neg\chi$ becomes a consequence of your information - in which case $\vdash$ is definitively non-monotonic. More modestly you might attempt to block the derivation of χ by removing some premises from our theory. Or perhaps reason indirectly to the conclusion $\neg\chi$ by the sheer absurdity of the conclusion χ and its consequences. For this latter strategy to work we would have to show that $\chi \vdash \bot$ or more likely T$\cup\chi \vdash \bot$. In any case an abduction problem is a choice problem.

A bit more precisely. Let an abductive problem be a χ-problem where we wish to conclude χ, but it is not immediately derivable from our theory T. So for instance we have a choice between the a set of explanatory or justificatory claims (γ, γ_1, ...γ_n). This set may or may not contain mutually exclusive elements. An abductive solution to the χ-problem is a choice function over $\{\gamma$, γ_1, ...$\gamma_n\}$ with one free variable i.e. c($\underline{?}$ $\{\gamma, \gamma_1, ...\gamma_n\}$) = γ_i such that $\gamma_i : \chi$ for one particular choice which justifies χ. The choice function may be specified in such a way so as to be independent or dependent upon broader details of the theory T. In most cases, this choice function is best understood as an operation which takes inputs from our active theoretical commitments.

In any case we assume that there are operations we can perform on our theory so as to converge on the result T $\vdash \chi$. An abductive problem is as we described above, but an abductive solution is a sequence of operations which ensures that $\chi \in$ T. Obviously we can amend our theory with further propositional information or particular justification-operations as in justification logic. Let γ be either an operation performed on, or a premise included in, our theory. We make three classifications.[8] An abductive solution is

Consistent If T $\cup\{\gamma\} \nvdash \bot$

Fully Explanatory If $\gamma \vdash \chi \wedge \neg \bigvee(\gamma_1...\gamma_n)$.

[7] Thought of as abstraction from our beliefs and knowledge however these are cashed out.

[8] Following work in [84]

Minimal If for all abductive solutions γ’ if $\gamma \vdash \gamma$’ then γ’ $\vdash \gamma$.

These criteria are apt if not exhaustive constraints on abductive solutions. Recall the Gettier case involving *Moriarty’s wallet.* It is clear that the argument works by undermining the explanatory power of any justification you might endorse. No typical justification γ can be such that T $\cup\{\gamma\} \nvdash \chi$, where χ is the Moriarty-hypothesis. For any typical scenario, there is an extreme counter-example to be found in logical space. As such Gettier concludes that justified true belief, no matter the particular nature of the justification, is not definitive or even explanatory, hence not sufficient for our understanding of knowledge. In short no addition of vindicating premises can exclude an underdetermination problem if we allow that W is the entirety of logically possible space. Even restricting W often allows for the generation of a reasonable underdetermination problem. This is, despite appearances, a positive result. Such an observation is crucial for underwriting the non-monotonic species of explanation.

The idea that a justification or an explanation must be such that it rules out all alternatives is too strong. This confusion stems from the idea that knowledge is that which holds in possible worlds, and explanations and justifications are apt to induce knowledge. This constraint insists that an explanation would ensure that no $\neg\chi$-worlds are possible, after an explanation has been deployed. But this only really holds if our justification is permissible in every possible world. Or put another way, the function c($\underline{?}$, $\{\gamma, \gamma_1, ...\gamma_n\}$) = γ_i such that $\gamma_i : \chi$ where γ_i is consistent, fully explanatory and minimal is very hard to find. Sceptical arguments go towards separating the notions of explanatory and justificatory claims from knowledge. We might prefer to say that an explanation (or justification) is apt to induce knowledge or belief just when it suffices to rule all other similarly plausible hypotheses. Construed in terms of plausibility orders we might say that an abductive solution to the χ-problem is an explanation i.e. an operation which re-orders the plausibility ranking so that only the χ-worlds appear among the most plausible worlds/theories. Or rather we might simply adopt the following standard: An abductive solution γ is...

Explanatory if $T \cup \gamma \vdash \chi$

Fully T-Explanatory if $T \cup \gamma \vdash \chi \wedge \neg \bigvee(\gamma_1 ... \gamma_n)$.

Preferred if γ is **Explanatory** and for all γ’, we have a ranking such that γ is preferred to γ’

On this setting we have the agent consider an abductive solution solely with respect to his total information, but we might further want to say that only true information is considered. What reason do we have to think that grue advocate cannot derive his desired conclusion from his total information. To avoid such possibilities we should have to specify further constraints on the calibre of an abductive solution. Minimally, we would hope to converge collectively on a single solution.

Minimal constraints on a Solution

You might think that abductive solutions to a χ-problem are primarily subjective. They are the operations required to enhance a theory in a way which ensures the derivability of χ from the newly available premises. But this “solution” is no solution at all. Add a contradiction to your theory and we can derive any number of trivial abductive solutions to a χ-problem. Easier still, add χ to your theory and we’re done. To avoid trivial problems we must impose a restraint against such solutions.

We say that a theory T *commands assent* just when true information regarding the theory is such that only one conclusion will be inferred by any rational agent in the χ-abduction problem. Allow that the notion of rationality is here somewhat context dependent, and we insist that rational reflection on theory T is performed with acknowledgement of the appropriate context. Contexts are incorporated in our reasoning by the addition of information and rules of procedural inference to our base beliefs. If T *commands assent*, then disagreement over the χ-problem is only possible if two agents reason on the basis of distinct assumptions about the context i.e. that they are working with relevantly different theories T_1 and T_2. We call reasoning from inappropriate contextual assumptions to be defective reasoning.[9]

An abductive solution is intersubjectively rational just when it is agreed that our theory is (a) contextually appropriate and (b) commands assent. If we let γ be the conjunction of all the contextually appropriate information and rules of justification, γ can be added to our theory. So we say that γ is an inter-subjectively rational solution to the χ-problem if $T \cup \{\gamma\} \vdash \chi$. This is far preferable to the merely subjective solutions to the χ-problem, since the intersubjective constraint forces us to defend our solutions. In this respect, there is an effective check on the viability of our abductive solution.

Any hope of putting an entirely objective check on candidate solutions flounders if we link objectivity with knowledge, due to the pervasive nature of Gettier-style counter examples. But in analogy with the notion of safe belief, we might think of an abductive solution as being objectively confirmed if under any true change of the context our solution χ remains derivable. When we discuss an abductive solution we would hope that it at least be intersubjectively rational. Ultimately, we wish to be able to say how the "?" input of our choice function, $c(\underline{?}\ \{\gamma, \gamma_1, ...\gamma_n\}) = \gamma_i$ such that $\gamma_i : \chi$, is determined.

4.2.3 The Rule Following Objection

The constraint of intersubjective rationality minimally means that no theory can be updated with an abductive solution without communal corroboration i.e agreement and acceptance of the "?" parameter. But there remains a problem regarding whether we can ever achieve this standard. Suppose we add a rule "?" to our theory, so that we have $T \cup \{\gamma\}$, now if our notion of justification is developed a la the semantics for justification logic in the last chapter, then we need to specify an evidence function $\mathcal{E}$ which describes the protocol for deploying our particular justification γ. The problem which crops up is that there is no constraint on how this protocol is followed. This is a version of Wittgenstein's Rule-following paradox.

Suppose you update your theory $T \cup \{\gamma\}$, and then seek to deploy the new justification to infer ϕ. You may always encounter a deviant logician who subscribes to the view γ: $\neg\phi$ since for any standard used to determine $\mathcal{E}$, the deviant logician can supply another function $\mathcal{E}^*$ which would motivate the opposite conclusion. For example, consider the series 0, 1, 2, 3... call it the natural series. When asked to enumerate the natural series the deviant may subscribe to the view that the justification for development of the series involves the application the +1-rule, but nevertheless diverge at n+1. So let ϕ be the statement that the n+1^{th} element in the series is the number n+1. We can define an evidence function $\mathcal{E}^*$, so that there is a justification for the conclusion $\neg\phi$, when the deviant's elaboration of the natural series diverges from ours. The only issue is to find a justification, but since plausibility need not be considered, this is alway possible. For instance we might think that the deviant's series reverts to the norm at the n+2^{th}-step, and continues as if to mask a hiccup.

To avoid such problems we might simply insist that our context precludes inordinate deviancy. We can do this in two ways: (1) we simply advocate against deviants, or (2) we observe that deviancy of such a stripe is incomprehensible or incredulous given our context. The former option is often question-begging on the occasion when the deviant's series is entirely consistent. However if we utilise (2) then we effectively privilege our current and best information to negatively rank, or preclude the deviant hypothesis. This is Glymour's bootstrapping strategy, since it involves

[9] A discussion of similar constraints takes place in [96]

the procedure of deriving our justifications from within our operative theory. In particular, we preclude certain species of procedural justification if they contradict the already utilised justification procedures e.g. the +1-rule cannot be used in a way which would disrupt the uniformity of the procedure due to the inherent meaning of the rule. We prioritise the available evidence over consistent alternatives unless further virtues can be found in the alternative hypothesis.[10]

Of course the rule following problem is a special case of an abductive problem, since it repeats for any solution you care to suggest. *It's a revenge paradox.* Whatever rule you propose as an input for our choice function to some particular abductive problem, we can generate an abductive problem relating to the use of that rule - in particular, we can challenge its proposed role in the choice function as a solution to our initial abductive problem. However, much the same strategy applies at each step. For each candidate application of a rule, we seek the best support for an application within our broader theory. Infinitely iterated stages of underdetermination are possible, but unlikely due to the discovery of inviolable procedures or standards within our broader theory. Stronger still these implicit rules or standards often form the bedrock of our theory of the world. Not only are they widely motivated by their central and profound role in our theory, but often our theory could not exist without these posits. They are well motivated because they can be discovered to be indispensable rote fixtures of our day to day.

Implicit Solutions and their Motivation

The rule-following problem like the grue-problem before it, offers the suggestion that some solutions to an abductive problem are question begging. We suggest that we may discover non-question begging rules of inference implict in our active theory such that these derived rules of inference can serve as a solution to the corresponding abductive problem.

The idea here is that we must seek within our active theory a reason to adopt a rule of inference. So for instance, take your best causal theory. The idea here is that our best causal theory of the world serves as the "?" input in our choice function. If we find that an event χ is caused by the event γ, then we might wish to adopt an inference rule: If "γ", then infer "χ". Let $T^* = T \cup \{\gamma \text{ causes } \chi\}$. We do not need to identify causal connection with logical entailment, but simply insist that causal patterns should have a role in our abductive reasoning. Including such a rule in our theory allows for the fact "γ" is abductive solution to the χ-problem. Since $T^* \cup \{\gamma\} \vdash \chi$.

As in the case of justification logic we need to inaugurate a categorical distinction between factual propositional information and the rules of inference included explicitly in our active theory. We can decide to adopt a rule independently of establishing it's truth, but this is not optimal for converging on inter-subjective agreement, never mind the more idealistic goals of sound reasoning. As such we should minimally establish the empirical adequacy of the proposed rule of inference. In our example above, this means that we should check whether the event γ does in fact cause χ. This suggests the irresistible conclusion that our best reasoning must explicitly incorporate the record of our causal metaphysics.

This option allows us to somewhat mitigate the accusation of circularity, since we now can see how an inference rule is motivated by our theory of causality. So for instance, instead of accepting the viability of the contention that God caused some particular tornado, we can adopt rules which circumscribe the actions of tornadoes in terms of the sole influence of the local weather systems. The adoption of these rules, being well motivated, allows us to apply them to exclude the God-cause hypothesis. Similarly we can conclude the falsity of grue hypothesis on metaphysical grounds only if we incorporate such reasoning by way of the appropriate inference rules. We can

[10] In [83] we find a dressed up notion of this kind of argument in Sider's appeal to "reference magnetism", the point is virtually the same. A more detailed discussion of the particular issue can be found in his essay "Ontological Realism" in [21]

challenge the mathematical deviant, if we insist that their application of the + operation cannot consistently lead to their conclusion. So long as we seek to motivate such rules from relevant sources independently of their application to the particular abductive problem, this is a legitimate solution to such abductive problems.

However there is an open question regarding the status of the adopted rules of inference. We might want, after a certain period of reliable usage to simply incorporate them in our theory without flagging the defeasible nature of the rule. At which point do we replace "γ causes χ" with "$\gamma \rightarrow \chi$", or do we ever? You might want to think that an adopted rule remains an adopted defeasible rule of our theory until the point where our theory is recognised as one which *commands assent*. In other words, if we achieve communal consensus over the truth of a theory which has incorporated defeasible rules we might begin to conflate those rules with laws of deductive implication. This is often a cause of error, but at no point do we say that this will necessarily be a cause of error. It all depends on the nature of the metaphysical claim that motivated our adoption of the inferential rule. Famously causality and material implication diverge, but this need not be true in the case of metaphysical dependence and material implication. Reliable usage and communal consensus contribute to more or less embed a given rule in our operative theory. The speed with which a rule becomes embedded will depend on the nature of its motivation, the expectation of reliability based on the underlying metaphysics and the robustness of our defense of this rule in a Lehrer-style justification game.[11]

4.3 Epistemology and Structural Information

We have been arguing for the idea that abductive problems can be stated but poorly solved with the standard tools of formal epistemology. We suggest that this follows from the inability to factor directly for structural information in our reasoning. An effort to include such structural information about the intrinsic connection between our premises and our conclusions (e.g. the colour of emeralds, our available information and the causal influence of the former on the latter). Only by including such structural information, by way of the adoption of certain rules of inference can we properly account for the manner in which, for instance, causal, modal, or mereological information can be used to develop a solution to an abductive problem.

This form of solution generalises to any number of cases of abductive reasoning. The challenge is only on deciding what kind of structural information is appropriate for which abductive problem. Ted Sider develops this line of thinking to the conclusion that "structure fills a need in epistemology".[12] An objection to this line of thought is to say that all talk of structure is determined by convention, so the use of such structural information is no more principled that that of drafting an arbitrary abductive solution. Sider asks us to consider notions of the geometry of space time - specifically he asks us to consider notion that space-time lacked a distinguished metrical structure, but that we have a successful joint-carving topology of space time. Then no metric would be distinguished one from another, and space time would effectively become an "amorphous point soup". As such...

> Reality might at the same time have lacked sufficient structure to define forces. In such a Reichenbachian world, we would have been free to choose either of a pair of coordinative definitions, simultaneously defining force and metric predicates. Neither choice would have carved reality at its joints better than the other. Metric and force predicates would require coordinative definitions in such a world, not because of general semantic considerations, but rather because the world would lack the structure needed to supply semantic determinacy. What reason do we have to think that our world has

[11] See [61] for further discussion

[12] [83]pg45

any more structure? The fact that physical theories with primitive metrical predicates have been so successful.

The argument here is in an effect a *reductio* against the notion of conventional structuralism. If we allow that structure is a purely conventional notion we undermine the adoption of our best scientific theories. From this we can observe that the success of our best scientific theories is paired with the implicit assumption of robust structural information. This argument may be questioned on two fronts, (1) if the success of a theory is based on the degree to which approximates the truth, then any theory which incorporates true structural information will be successful. But then our claim that structural information is vital for success begs the question, (2) if the success of the theory is based on pragmatic considerations then our argument is sound but irrelevant since we have not established that we need anything more than some suitably effective assumptions of a structural kind. Suppose success comes from a combination of predictive power, simplicity and explanatory breadth. Then our structural assumptions are not inherently pragmatic because they allow us to predict particular truths, thereby increasing the degree to which our theory approximates truth. Similarly, as we shall show below structural assumptions lend a theory explanatory breadth, and, so their adoption does not simply beg the question that a successful theory is successful, because it contains true structural information.[13]

This bears on the epistemology of metaphysics. Sider argues for the broadly Quinean contention that we should take the ideology of our best theory as carving at the joints of reality because this gives us defeasible reasons for taking our best theory to be (minimally) indicative of reality's structure.[14]. The fact that the theory is the best available implies it has a broad motivation independent of its application to our candidate abductive problem. Hence our rejection of structural conventionalism with regard to space-time is well justified.Sider himself makes a much stronger claim that the notion of a distinguished ontological structure is inextricably bound up in our understanding the world: "The very idea of distinguished ontological structure itself, once grasped, is one that must surely be acknowledged."[15] With such a strong conclusion, you might expect him to deploy a charge of incoherence against the conventionalist. None is forthcoming. However, we might argue as follows: structure is inherent to our conceptualisation of the world, we take the world to be the ultimate arbiter and measure of truth and falsity. To accept that conventionalism is true minimally entails your acceptance of two mutually exclusive theories T_1 and T_2 as being equally empirically adequate forever. This undermines the idea that the world can serve as the ultimate arbiter and measure of truth and falsity, since the world is consistent. Hence the conventionalist must be working with an incoherent notion of truth, or at least must defend the idea that the choice between T_1 and T_2 is forever underdetermined - but lacking an argument to this end the conventionalist begs the question against Sider. In any case, we take it that a world without structure is no world at all.[16] We turn now to the question of how such structural information can help resolve the issue of perennial underdetermination in epistemology.

4.3.1 Epistemology of Mathematics: A Hard Case

We now turn to a famous case of underdetermination. If what we have already argued holds we should be able to resolve this underdetermination problem by incorporating a well motivated abductive solution. We aim to show how incorporating structural information facilitates this type of move, and thereby plays a role in epistemology even in the difficult setting of the philosophy of mathematics.

[13] Sider himself cites as fruitful examples, Frege's insistence on quantificational structures in reasoning, and Chomsky's insistence on respecting the non-prescriptive judgements of native speakers regarding grammatical structure. cf. [21]pg402.

[14] [21] and [83]pg14-18

[15] [21]pg402.

[16] We offer a more full throated defence of this structural realism in our final chapter.

Recall Benacceraf's famous parable of the militant set theorists.[17] The idea was that under certain conditions of domestication we could evolve a breed of mathematician who would work as naturally with a set theoretic reduction of the mathematics, as we currently operate with the unanalysed notions of number, function and operation. This is presented as a problem for the structuralist analysis of mathematical ontology, since it turns out that not only is our first experiment in social evolution successful, but that we can do the same with a rival set theoretic reduction. Hence the identification of the pre-theoretic notion of number, with the notion of a set theoretic construction is ambiguous between two distinct species of set theoretic constructions.

Only upon meeting one another do our two trained set theorists realise their problem. Call one Ernie, and the other Johnny. Suppose Ernie learns the Zermelo construction of the number series where, the successor of any set $n = \{n\}$ and Johnny learns of the Von Neumann construction where the successor of any set $n = \mathrm{n} \cup \{n\}$. In both cases their ultimate constructions are adequate to function as formal analogues of the intuitive number series. Worse news for Ernie and Johnny, they cannot even speak to one another without the presence of a multi-lingual mathematician. Their respective use of the union operation is incomprehensible to the other. However, since both languages involve fundamentally different constructions, we feel that the intuitive series cannot be properly identified with both constructions.

> If numbers are sets, then they must be particular sets, for each set is some particular set.[18]

However due to the nature of the Zermelo construction we know that since 1 is identified with the set $\{\emptyset\}$ that $1 \in 17$, but on the Von Neumann construction $17 = 16 \cup \{16\}$, and so $1 \notin 17$ according to the Von Neumann construction. Yet both constructions serve adequately to derive the formal properties we have come to expect of the number series.[19] Hence, each proposed reduction of the intuitive notion of numbers to a set theoretic construct is fundamentally underdetermined and we are faced with a problem of arbitrary identifications if we choose either way. Seen as an abductive problem we might wonder how to solve this problem? What would count as an abductive solution to a question in the ontology of mathematics?

In this case we suggest that the appropriate line of thought is to consider in detail the notion of the grounding relation, and more broadly our theory of metaphysical dependence. Since we seek to run a reduction of numbers to one or other set theoretic construction, we should consult our theory of grounding to determine which (if any) is the more appropriate candidate. On the face of it, each reduction is equally useful for serving as a foundation of mathematics, but we have said nothing of which reduction (if any) represents the more plausible metaphysics.

We follow Benacceraf and choose to make this a metaphysical issue because we feel it is the only appropriate course of action. Potentially you could also see this as an issue of cogent definition, and so seek to bar one or other set theoretic reduction on the basis of the semantically incomprehensible notion of membership entailed by one of the candidate constructions. I contend that such a course addresses the same problem as we wish to, but simply stipulates the correct semantics of membership whereas you should alternatively seek to motivate the notion of mereological composition which underlies our semantic intuitions regarding membership. In a sense, any abductive solution sought on the basis of purely semantic considerations will be derivative upon the properly metaphysical solution.[20] If, that is, there is one to be found.

[17] For details see [13] and [14]

[18] [13]

[19] That said, Steinhart has recently argued for the pragmatic reasons to prefer the Von Neumann construction in [89]

[20] If we hope to maintain a tolerable truth-conditional semantics for mathematical statements

The Grounding Relation

The notion of a grounding relation is a philosophical term of art which is advertised as having some simple intuitive instances wherever we say that ϕ holds *in virtue of* ψ. Clark and Liggins argue that grounding is closely related to notions of explanation; and they cite a number of reasonably intuitive cases where grounding relations track a hierarchy of more to less fundamental ontological features of the world, where the fundamental features are necessary condition for the existence of the derivative (or complex) conditions.[21] That is to say, grounding relations track a non-causal form of dependence. Consider the following examples.

(E1) The dispositions of a thing are always grounded in its categorical features. For example a cup is brittle because anything which is a cup supervenes a more fundamental arrangement of molecules, and the laws of physics and chemistry which govern the current arrangement of the molecules allow exceptions. Hence the dispositions of our cup range between solid state and broken, as fully determined by the flex of the categorical laws governing the cup's constituent elements.

(E2) The existence of legislation is grounded in broader institutional and social facts primarily relating to the enforcement or acceptance of the law.

Despite some recent protestations[22] the grounding relation has been traditionally thought of as irreflexive, asymmetric and transitive. That is to say, it is a strict partial order. Furthermore Cameron has recently argued apart from theoretical nicety, that there is little reason to think that the grounding relation is well-founded.[23] For to think that there is a fundamental ontological level containing only mereological atoms is an empirical belief not yet motivated by the progress of science.

Schaffer argues similarly that the reductive projects which would make higher order entities epiphenomenal or causally inert, are based on the misleading presumption that our base-information is always well founded. So to call some or other phenomena epiphenomenal is often more an indication of poor understanding of the phenomena relative to the understanding of our base assumptions, than it is a well motivated metaphysical analysis.[24]. If no ontological level is well founded, then all are composite and all ontological levels are equally apt to be the loci of causal powers. If we allow that grounding is a non-well founded relation we can quickly distinguish it from a number of candidate relations in the area.

Significantly, Kit Fine insists that the grounding relation should ignore any kind of monotonicity restraint. Since we might want to say that ϕ holds in virtue of ψ, but it would be inaccurate to say that ϕ holds in virtue of $\psi \wedge \tau$ where τ added nothing to the preconditions for ϕ given ψ. For of course, if we were to insist that monotonicity held, then every fact could be stated to have a role in the grounding of every other. This is simply false, but it is made doubly clear if we distinguish between full and partial grounding relations.

The notion of a weak grounding relation is opposed to the stricter irreflexive grounding relation of strict grounding. Similarly the notion of full grounding relation is opposed to the notion of a

[21] [23]

[22] For instance, cf. [51] where Carrie Jenkins argues that wholes may be though of as being grounded in the sum of their parts. She argues, for instance that a statue is grounded in the sum of it's parts, but also is equal to the sum of it's parts. Hence grounding can be thought of as a reflexive relation. Similarly, Schaffer has argued that there are counterexamples against the transitivity of ground. Consider the fact that something is meowing, observe that it is grounded in the fact that Cadmus the cat is meowing. While Cadmus meowing exists because of the copulation of his parents. Yet we cannot say the fact that something is meowing is grounded in the fact of Cadmus's parents and their copulation.

[23] [18]

[24] [81]

partial grounding relation where the latter relation intuitively holds between a collection of facts Γ and some particular fact ψ, where none amongst the collection of facts Γ is sufficient to individually ground the fact ψ. Naturally a full grounding relation holds only amongst facts where the ground is the sole explanatory condition for the truth of the emergent phenomena.[25]. Raven[26] suggests that Γ partially ground ϕ just when there is some p $\in$ ϕ such that Γ fully grounds p. The non-monotonicity of a full grounding relation is self evident. A counter example to the monotonicity of partial grounding involves the addition of grounding facts until the grounds exceed the information determined by the fact of emergent phenomena i.e. Suppose ϕ is partially grounded in ψ, now add p_1, p_2....etc to ψ to the point where $\psi+$ becomes fully grounded in ϕ.

4.3.2 The Logic of Full Ground

The language of the logic of full weak ground is surprisingly simple. Most of the complexity rests in the definition of the model. We construct $\mathcal{L}_{pg}$ as usual.

$$\phi ::= p \mid \phi \precsim \psi \mid \Delta \precsim \psi \mid \phi \prec \psi \mid \Delta \prec \psi$$

Where p_i is in the set, possibly infinite, of atomic formulas, and we represent statements of the form "ψ is weakly and fully grounded in ϕ" as $\phi \precsim \psi$. While Δ is a set of atomic statements, and the claim $\Delta \precsim \psi$ just means that ψ is fully and weakly grounded in the set of truths Δ. Let $\precsim_\pi$ be ambiguous between $\precsim$ and $\prec$ where $\prec$ is the strict form of the full grounding relation. We have the following rules of inference for a sequent calculus.

Subsumption

$$\frac{\vdash \ \Delta \prec C}{\vdash \ \Delta \precsim C}$$

Cut

$$\frac{\vdash \ \Delta_1 \precsim \phi_1, \Delta_2 \precsim \phi_2 .... \phi_1, \phi_2 \precsim \psi}{\vdash \ \Delta_1, \Delta_2 ... \precsim_\pi \psi}$$

Identity

$$\frac{\vdash}{\vdash \ \phi \precsim \phi}$$

Non-Circularity

$$\frac{\vdash \ \phi, \Delta \prec \phi}{\vdash \ \bot}$$

Reverse Subsumption

$$\frac{\vdash \ \phi_1, \phi_2 ... \precsim \psi, \ ((\phi_1, \Delta_1) \prec \tau_1), \ ((\tau_1, \Gamma_1) \precsim \psi) \text{ etc, etc for all } \phi_i, \Delta_i, \tau_i, \Gamma_i}{\vdash \ \phi_1, \phi_2, \prec \psi}$$

[25] See [28] for further details and a discussion of the relevant distinctions

[26] [75]

The reverse subsumption rule is an ugly formulation of a rule which is more easily defined using the notion of a partial ground. It is supposed to express the thought that that where a set of $\phi_1 - \phi_n$ is a weak full ground for the truth of ψ, and for all ϕ_i, we can be show ϕ_i to be included in a strict *partial* ground for ψ we can infer that the collection $\phi_1 - \phi_n$ is also a strict full ground for ψ. Think of the case where a parliament is weakly fully grounded in the existence of its members, so each member of parliament is a partial ground for the existence of parliament, hence we conclude that the set of members of parliament strictly fully grounds the existence of parliament. The other rules are more intuitive, for instance Cut preserves the transitivity of the grounding relation, while Non-Circularity ensures that strict full grounding is an irreflexive relation. However Identity allows for the possibility that Jenkins' heresies about the reflexivity of the grounding relation are not without merit. Finally, the non-monotonicity of the grounding relation is ensured by the absence of a Weakening rule. So far so good, what about the semantics?

The Semantics

Consider the notion of a verification set as follows: think of $[\phi]$ as the *verification set* for the truth ϕ, perhaps construed as the set of facts at the actual world sufficient to ensure the truth of ϕ. That is to say that we need not think of a truth-making relation as of necessitation, but instead as the relation of "relevantly verifying". Kit Fine suggest this treatment with the proviso that "we should think of the facts of $[\phi]$ not as *possible* worlds, but as *parts of the actual world*"[27] The verification set $[\phi]$ can be seen to contain a number of facts, all of which might obtain in the world simultaneously. So suppose *f, g, h* are all and only the facts sufficient to verify the truth of ϕ, then we may suppose that if they obtain, in the actual world, then the verification set $[\phi]$ is just the mereological fusion, or composite fact *f·g·h.* With this definition in mind Fine proceeds to define a semantics for the logic of various grounding relations.[28]

Fact Frames We let the following ordered pair be the skeleton of our semantic model. A frame $\mathcal{F} = < F, \Pi >$. Where F is the non-empty set of facts (read states of affairs) and Π is function, (the factual fusion) which takes each subset E of F into the factual fusion $\Pi(E)$ where naturally the fusion of all component parts of the subsets of E, is identical to the fusions of the fusions of the subsets of E. Expressed more symbolically we have:

$$\Pi\{\Pi(E_i \mid i \in I\} = \Pi \cup \{E_i \mid i \in I\}$$

Hence we expect two limit points where $\Pi(E)$ is defined for E = $\emptyset$ and E = F to exist. Of course, when E = F, we know that $\Pi(E)$ = w, i.e. the world w, the actual world. It's less clear what role the null-fact fusion will have to play. Although, if we admit the existence of ontological atoms, we could treat them as being grounded in the null-fact. In any case, if we wish to insist that the full grounding relation is upwardly closed we should expect any truths relevantly verified by the null fact, to be similarly verified by the null-fact fusion. We call the null-fact fusion $\mathrm{n}_{\mathcal{F}}$ and the full fact fusion $\mathrm{w}_{\mathcal{F}}$.

$$\text{E} \subseteq \text{F is upward closed if } \forall D \in E(D \text{ is non-empty} \rightarrow \Pi(D) \in E)$$

This condition is adequate to ensure that the truth of any particular set of facts holds throughout any additional fusions of said facts. By specifying that D is non-empty we avoid any issues of having to determine what facts if any the null-fact fusion would verify. Crucially of course, all facts are grounded in the full fact Π(F).

[27] cf. [28]pg7

[28] We simplify the presentation of this logic as for current purposes this is purely illustrative.

Fact Models A fact model is a fact frame with a function which stipulates which facts actually do serve to relevantly verify (ground) our particular claims. We denote our fact model M = $< F, \Pi, [\,] >$ where [ ] is a function that takes each atom p_i of our language into a non-empty closed subset $[\phi]$ of F. Intuitively this is supposed to determine those facts in F which actually would verify our claims. On the assumption here that every true claim will have a verifying fact, then we know that there is a non-empty verification state $[\phi]$ for all claims ϕ. This is the constraint our function [ ] is supposed to enforce. The function in effect constrains the type of features of the world which *would* serve to verify our claim ϕ.

Generalised Fact Frames In addition to being a fact frame a generalised fact frame comes with a function which stipulates which subsets of F, do in actuality serve as grounds for other subsets of F. So $\mathcal{F}_{gen}$ is a generalised fact frame when $\mathcal{F}_{gen} = < F, \Pi, \mathrm{Ver} >$ where Ver: $\{(x \mid x \in \upharpoonright \wp(F)\} \mapsto [True]$. We say that Ver is full if it takes all non-empty closed subsets of F to the truth. That is to say, Ver is a function which determines those facts which can, in virtue obtaining in actuality, serve as grounds for the verification of our claims. The restriction determined by Ver can be seen as being enforced by the limitation of observation, or more generally our evidence gathering procedures. It specifies our verification space, and if Ver is full i.e. takes all subsets of the state space F, then the entire world is the verification space. Optimistic perhaps, but no metaphysician should balk at the prospect.

Generalised Fact Models Putting all the foregoing considerations together we get $\mathcal{M}_{gen}$ = $< F, \Pi, [\,], \mathrm{Ver} >$ where F, Π and Ver are as before, and [ ] is defined to send atomic sentences into dom(Ver). Which is to say, that we first determine our verification space and then assess the truth of our claims with respect to the available facts. Hence [ ] is a kind of valuation function conditional on the state of the world as determined by Ver.

4.3.3 Relevantly Verified?

The reason why we can say that ϕ grounds (or relevantly verifies) ψ, is because we know that that the facts which verify ψ already contain the grounds for the verification of ϕ and these facts have been determined to be true at the actual world. This is a metaphysical claim and our metaphysics is given by Ver and [ ] operating in conjunction. As such the idea is to say that the metaphysical structure of the world determines the relevance of one claim to another, and we have already adjudged the world to have a particular grounding structure. This will be clearer after we observe the satisfaction clauses for our language. Let G be a possibly infinite set of fact $\{f_1 ... f_n\}$.

Definitions: Truth

(Containment) $\mathcal{M} \models \{f_1, f_2 ... f_m\} \precsim$ G if $f_1 \cdot f_2$ .... $\subseteq$ G.

(Satisfaction) $\mathcal{M} \models \Delta \precsim \psi$
$\Leftrightarrow [\Delta] \precsim [\psi] \in \mathcal{M}$
$\Leftrightarrow \Pi[\Delta] \subseteq [\psi] \in \mathcal{M}$
$\Leftrightarrow \forall \mathrm{p}_i \in \Delta$ (if $\mathrm{f}_i \in [p_i]$, then $\Pi(f_i) \in [p_i] \in [\psi] \in \mathcal{M}$[29]

Which in effect says that ψ is grounded in Δ just when ψ entails Δ, but more importantly the fact of $[\psi]$ is grounded in the fact $[\Delta]$, and we can "read off" this structure relation from the metaphysics determined by V from the final clause of the satisfaction definition.

You might wonder at this point why - recalling that the rule following objection was the problem that stemmed from the difficulty of specifying conditions under which a conclusion was relevantly

[29] A completeness result via a canonical model construction is developed in [28] so any consistent derivable sequence of the Pure logic of ground has a model i.e. with the logic with four sentential connectives is sound and complete. As a fragment of this logic, a completeness theorem for the pure logic of full ground is recovered.

linked by a rule to certain premises - we should think that the notion of relevant verification helps solve the problem? Surely we've just stipulated the solution by incorporating rules of relevance? This is to miss the point altogether.

The point. If grounding claims are a species of explanatory claims, and we can use our metaphysical theory to motivate explanatory claims, then we should always try to locate our explanatory claims within a broader metaphysical setting. Relevance is determined by dependency relations, we have not simply defined relevance in terms of such dependency - it already was constrained by dependency information! The moral here is that we should first fix our metaphysics, before attempting to find any abductive solution. With this structure in mind, we seem to be in a good position to develop a picture of explanatory claims as those which track dependence relations amongst the facts of the world. Let T$_{pg}$ be our theory of the grounding relation, then for any χ-problem such as our hard case we define a choice function c($\underline{T_{pg}}$, $\{\gamma, \gamma_1 ... \gamma_n\}$) $= \gamma_i$ where $\gamma_i : \chi$. Then the properties of the explanation for χ will be derivative upon the properties known to hold of our grounding relation.

4.3.4 Our Hard Case Reconsidered

Suppose that the reductive projects of our set theorists succeed in so far as they do because they have observed that the individual numbers are grounded in the sum of their parts. First we should note that this case is analogous to the notion that a statue is equal to the sum of its parts, and so consequently to achieve the metaphysical identification of individual numbers with a particular collection of sets, we should first have to adopt a reflexive relation of grounding. This remark holds for both set theoretic constructions.

As such we could argue that both set theoretic reductions are illegitimate because there is no sensible notion of a reflexive grounding relation. So neither proposed reduction can serve to replace our intuitive notion of number. Or worse, we could conclude that given all we know about the grounding relation we can see that there is no role for grounding in the debate over the existence of numbers. Grounding is relation of wholes to observable parts, and the contention that numbers have parts (of any description) is question begging in favour of some or other set theoretic reduction if the description is given in set theoretic terms.

If we do accept the notion of a reflexive grounding relation than metaphysically there is no preference between either construction, in the same way that for all intensive purposes there is no difference between a coke bottle constructed in a factory, and one made for display in a post-modern protest against consumerism. They both hold water. This is not a renunciation of metaphysics for pragmatic considerations. The idea is that both architects are already committed to the existence of numbers, in the same way that Rodin is committed to the abstract ideas of thinking. The mere fact that they seek to identify particular constructions with abstract idea does nothing to ensure that their identification is any more than convenient, or simply evocative.

Similarly if we follow Schaffer and argue for the non-wellfoundedness of the grounding relation, then the attempt to define particular numbers in terms of the well order that is the set theoretic construction, becomes immediately bankrupt or (again) question begging by the presumption that numbers have parts (and indeed foundational parts). Although, this may be more of a flaw in Schaffer's position than it is a reason to doubt the set theoretic reduction, as this position would seem to force us to divorce our understanding of numbers from our understanding of the corresponding set theoretic constructions altogether! Perhaps Schaffer takes the reals as primitive? This again would seem to dissolve the problem rather than solve it. Alternatively if we take the notion of a full and partial ground seriously, we might say that for any number n, we know that n is fully grounded in the Zermelo construction and only partially grounded in the Von Neumann construction due to the fact that the latter construction is only partially explicit about the sets required to ensure the existence of the number n as opposed to the forthright presentation of the Zermelo dependencies. But this strikes me as a very flimsy motivation.

Our hard problem can be seen as an abductive problem only if one of the options is genuinely preferred, but since there is no reason to prefer either option, and no metaphysical information which would provide a clue as to the correct reduction, it's not an abductive problem at all. We might begin to think that set theoretic reductions are not operative in the debate over mathematical ontology or that they miss the point altogether - why does mathematical ontology does require a reduction. Perhaps Plato was right after all, so at best set theory provides us with two workable heuristics but not a substitute ontology. But this undercuts Benacceraf's contention that if numbers are sets they must be particular sets, so the abductive problem dissolves. We shouldn't believe numbers reduce to either set theoretic construction and we certainly don't know that they do. They might be adequately characterised by set theoretic constructions, but this is a much different claim.

The discussion of metaphysical grounding has shown that the attempt to cash out our intuitive notion of number in terms of sets is at least as question-begging as the attempt to cash out our intuitive notion of number in terms of a particular type of sets over another. As such, we suggest that Benacceraf's underdetermination argument does not prompt an abductive problem at all since we do not have sufficient grounds to think this is even a reasonable choice.[30]

In any case the structural information incorporated by our examination of the grounding-facts allows us to develop a response to certain kinds of abductive problem, and indeed identify the insoluble abductive problems relative to the available (or agreed upon) information. In the above case we saw that either the abductive problem was ill-formed or under specified with respect to the operative notion of grounding. The crucial take-away from our discussion is that structural metaphysics has a real and unavoidable role in epistemology. Failure to incorporate such structural information makes a mockery of any formal epistemology in so far as it generates a plague of genuine Gettier-cases. *Knowledge becomes an unachievable goal, and belief is always at best question begging.*

4.4 Grounding and Explanation: A Model to Mimic

To alleviate this issue we shall have to explicitly link abductive solutions with epistemic or doxastic states. We consider briefly the idea proposed by Trogden.

First distinguish between full and partial ground. We claim that any full-grounding relation is a necessary relation. Which is just to say that necessarily, if ϕ is fully grounded in Γ, then it is metaphysically necessary that wherever $[\Gamma]$ occurs, so too does $[\phi]$. We now follow Kelly Trogdon[31] and argue for the latter claim by noting that (1) the notion of full-ground is an explanatory notion, and where any claim ϕ is said to be fully-grounded in Γ we are invoking the idea that it is absurd to doubt ϕ and accept Γ. Add to this that (2) wherever it is absurd to ask whether ϕ when you know Γ, there is some *essential* or modal connection between the fact $[\Gamma]$ and the occurrence of $[\phi]$. Finally, (3) any essential connection, is *ipso facto* a necessary connection. So suppose that you are offered an explanation for $[\phi]$ in virtue of the fact that ϕ is fully grounded in Γ, then by applying (1) and (2), you know that there is an essential connection between the fact $[\Gamma]$ and the occurrence of $[\phi]$ so by (3) you infer that $\Box(\Gamma \rightarrow \phi)$.

[30] This problem requires minimally that we develop a decisive argument against Platonism in the philosophy of mathematics. As it stands, the lack of such argument ensures that no abductive solution to the problem of multiple reductions can be properly speaking explanatory because there remains a possibility that (whatever our choice) our answer will not be entailed by our theory.

[31] cf. [54]

The essential connection between the two is, in the grounding case, supposed to track a metaphysical necessity. However, arguably the modal connection between the truth of ϕ and Γ need not be a metaphysical modal connection, we might for instance simply be tracking an *a priori* connection if we are merely tracking a relation of conceptual dependence. So if we say that p is fully grounded in the fact that (p $\wedge$ q), and (p $\vee$ q) is fully grounded in the fact that p, it's not clear whether it's absurd to question this connection, because the connection is metaphysically, or epistemically necessary. It's certainly logically necessary, so perhaps we should separate these species of explanation? In any case, the substantive claim here is that facts about whether or not one thing grounds another have an epistemic component! It is a fundamental marker of a full-grounding claim that minimally it induces the expectation and acceptance of ϕ, where Γ. So by analogous considerations we should insist on a similar although suitably qualified feature for claims of partial grounding.

In general we should expect explanatory claims to have a (1) epistemic and (2) factive component. Explanations are (we shall argue true) answers to why-questions, which determine a change in our epistemic (doxastic) state. In each case we can develop a similar argument, amending the epistemic component to suitably reflect the strength of the dependency relation. Perhaps allowing that certain abductive solutions only raise the credibility of one option over another. The correct characterisation of the effect of particular abductive solutions should depend entirely on the analysis of the appropriate kind of dependence relation. Whatever the nature of the particular dependency we may inaugurate a new update operation to encode such information in an agent's epistemic state.

4.4.1 Explanations and Expectation

Recall that most of our explanatory relations are non-monotonic, then note that if we wish to incorporate rules which track these relations we should adopt some terminology. We define as follows:

Default Rules $\phi \rightsquigarrow \psi$ iff ψ is true at all the most normal ϕ-worlds.

Non-monotonic consequence T $\vdash \phi$ iff ϕ is a consequence of our theory T$\cup\{\Gamma\} \vdash \phi$

The idea Boutilier[32] suggests is that we think of normality as a preference order on worlds. Specifically as the source of a plausibility-style ordering. In this way we ensure that the most plausible world is one which vindicates all our default rules. Minimally, you might argue that plausibility has to respect normality. By allowing that plausible worlds are dependent on a ranking of normality we will in the large part be able to solve any arising abductive problems and so avoid the state of paralysis they would otherwise prompt. Since the normality conditions give us access to a set of normative laws of inference, which can be used as abductive solutions in the cases of underdetermination.

Default rules can be thought of here are as an inference license, when one has been established it frees you make to make an abductive inference privileging the norm. In effect the establishment of a default can be seen as decision regarding what counts as a best explanation. The most significant objection to this line of thought is that we cannot insist that the normality ranking is primitive, for then the normality ordering is as question begging as an unanalysed plausibility ordering. However, it is not clear how to analyse normality.[33] But ignoring this difficulty for the moment, we wish to suggest that some operative notion of normality can be seen to underlie various species of explanation. We can immediately distinguish some categorical features of different types of

[32] [16]

[33] We will attempt to resolve this issue later by replacing default rules, with statements of evidentiary relevance. Instead of $\phi \rightsquigarrow \psi$, we can have a state j : ψ, where j := ϕ. In this manner we can bypass the problem of normality if we can suggest how evidentiary relevance is established.

explanations. The hope is that these categorisations are sufficiently general so that the adoption of a rule motivated by any particular type of structural dependency relation can be seen to induce only one of these categorical changes on our epistemic or doxastic state. Let's begin with the distinction between *factual* and *hypothetical* explanations.

Factual We observe the the truth of (or believe) ψ, and seek a fact ϕ which can be said to explain the occurrence of ψ

Hypothetical We do not know (or believe) ψ but seek to know or believe the conditions ϕ which would induce belief (or knowledge) in ψ.

So for instance, if we discover the grass to be wet we should seek a factual explanation for this occurrence, that is we must reason abductively to find a cause, or reason for the state of the grass. Discovery of this fact is apt to induce knowledge of (for instance) the recent bout of rain. Now we have choice, if we only believed that the grass was wet, our subsequent discovery that it has rained might be sufficient for us to claim knowledge that the grass was wet. If the appropriate *ceteris paribus* law holds in our theory T, then we might think that $\mathrm{T} \vdash \phi$, and since knowledge is typically closed under logical consequence, $\mathrm{T} \vdash \mathrm{K}\phi$.

In contrast the search for hypothetical explanations is based on purely conjectural reasoning. This does not make the explanations achieved any less integral to our reasoning. Counterfactual reasoning of that kind severely constrains our roving imagination. But again the issue of whether a hypothetical explanation should induce belief or knowledge is often dependent on the kind of hypothetical under consideration and the degree of reliability (if any) associated with such considerations. For instance, we might think that hypothetical causal reasoning ought to be trusted due to inevitable familiarity we have with causal processes and reasoning. But such an insistence is more dubious when we consider claims about grounding relations.

Predictive Explanations

We do need to link the notions of an explanatory (or default) rule with predictability if we wish to account for one of the most prevalent uses of abductive reasoning. Minimally we suspect that an explanation can only be predictive if belief in that explanation is sufficient to induce belief in the observation predicted.

Definition (Predictive Explanation) Let T be an epistemic doxastic theory reflecting the knowledge and belief of an agent and $\mathcal{L}_{\mathrm{T}}$ be the language of our theory. A *predictive explanation* for any observation ψ relative to T is any $\phi \in \mathcal{L}_{\mathrm{T}}$ such that

1. $\mathrm{T} \vdash (\mathrm{B}\phi \leftrightarrow \mathrm{B}\psi) \wedge (\mathrm{B}\neg\phi \leftrightarrow \mathrm{B}\neg\psi)$
2. $\mathrm{T} \vdash \phi \rightsquigarrow \psi$
3. $\mathrm{T} \vdash \neg\psi \rightsquigarrow \neg\phi$

The thought behind the three conditions can be unpacked as follows: (1) By the nature of most predictive explanations, they ought to be factual so where the fact ψ has been observed it is minimally believed and anything which would explain ψ ought also to be believed. In ideal circumstances this should hold. However our abductive problems show that there are counterexamples to the left conjunct. Similarly the right conjunct states that if ψ is not believed than anything sufficient to induce belief in ψ should also be denied. So evidently, a similar problem arises where the belief that $\neg\phi$ is not sufficient to induce belief in the falsity of ψ if there is another candidate explanation χ in the vicinity. The idea here is that all the maximally normal worlds would rule out such other candidate explanations. (2) This condition is crucial for predictive explanations as it allows to incorporate ψ in our theory T whenever $\phi \in$T. The relation $\rightsquigarrow$ can be thought of tracking the historical sequence, but to accurately test the claim we need, so to speak, to give up our observation of ψ, and consider only those most plausible counterfactual worlds where ϕ holds, and there observe the truth of ψ. Under such conditions we can say that ϕ is *ceteris paribus* a predictor for ψ.[34] (3) Our final condition is a hang over from the idealistic impetus for (1), it seems strictly speaking false, unless you wish ϕ to be the only explanation for ψ. This happens just when all other possibilities are quashed i.e. when ϕ is **fully explanatory** of ψ.

Non-Predictive Explanations

This species of explanation is much weaker the predictive strain of explanation. The crucial idea here is that we allow for competing explanations so that even where T $\vdash \phi \rightsquigarrow \psi \wedge \neg\phi$, we still may believe or know the truth of ψ where there is some $\chi \in$T such that T $\vdash \chi \rightsquigarrow \psi$. In other words there are a number of candidate answers for our why-question, but none of these answers are absolutely definitive. Boutilier calls this a species of might-explanations where:

> Intuitively, a might explanation reflects the slogan *If the explanation were believed the observation would be a possibility...*If an agent accepts explanation ϕ, the observation ψ becomes consistent with it's new belief set.[35]

Again we can avail of the notion of normality to define non-predictive explanations. On the assumption that non-predictive means that there will be a least one maximally normal ϕ world where ψ holds.

Definition (Non-predictive Explanation) Let T be an epistemic doxastic theory reflecting the knowledge and belief of an agent in $\mathcal{L}_{\mathcal{T}}$, the language of our theory. A non-predictive explanation for any observation ψ relative to T is any $\phi \in \mathcal{L}_{\mathcal{T}}$ such that:

1. T $\vdash$ B$\psi \rightarrow$Bϕ
2. T $\vdash \neg\ (\phi \rightsquigarrow \neg\psi)$

The idea behind the second condition is that ψ must be possible in all the normal ϕ-worlds. Intuitively, we want to consider some maximally plausible ϕ world where ψ is true. We need not insist that every maximally normal ϕ-world is also a ψ-world. It is guaranteed by the second condition that there is a maximally normal world where both ϕ and ψ hold. Of course, the second condition simply encodes the fact that ϕ is an abductive solution for ψ.We will return to the discussion of species of explanation later, but for now we use the foregoing to assess Benacerraf's dilemma.

4.4.2 Our Hard Case Again

Now if we apply the above distinctions to our considerations regarding the explanatory power of the two set theoretic reductions, we should see that neither proposed reduction is apt to be considered a predictive explanation. Both proposed reductions are deemed sufficient to establish the existence of numbers with their "normal" properties, so the falsity of either one will not ensure the non-existence of numbers, thereby contradicting the third condition in the definition of a predictive explanation. Similarly, while our belief in the existence of numbers might be thought sufficient to induce belief in one or other of the set theoretic reductions, the other direction fails. The belief in the existence of a set theoretic construction, is not alone sufficient to generate belief in the existence of numbers, without first presuming a belief in the fact that the numbers are emergent from a particular set theoretic construction. Seeing that the latter move is question begging ensures that the first condition in the definition of predictive explanation cannot be met.

[34] Boutilier in [16] articulates this idea with reference to the revision of our theory with ϕ. The details are not terrifically important but a nice summary and discussion of AGM belief revision theories can be found in [93].

[35] [16]pg22

So now let's consider whether either set theoretic reduction can be thought of as a hypothetical non-predictive explanation. Indeed they can, assume that Platonism in mathematics is false, and then either set theoretic reduction is an apt non-predictive explanation for the existence of numbers since numbers will be seen to have a constructive existence. Both constructions are possible since they both validate all the "normal" properties of the number sequence. Although strictly speaking, we're not even providing a normality claim; rather for each reduction we are claiming that the appropriate set theoretic construction serves as evidence for the existence of numbers. So instead of a default rule $\phi \rightsquigarrow \psi$ we should really state that there is evidentiary term e, which encodes ϕ, i.e. e_{ϕ}, and e_{ϕ}: ψ. But, in any case, such non-predictive explanations cannot be expected to rule out all alternatives. The abductive problem dissolves quickly in this case precisely because the explanatory task has been achieved as well it currently can. Non-predictive explanations do not allow us to achieve a unique inference to the *best* explanation. In retrospect this should have been obvious, the process of inference to the best explanation, could never uniquely determine a mathematical identity, by its very nature. It is simply not the task of a hypothetical non-predictive explanation to definitively establish the one and only cause of the event (or phenomenon) for which we sought an explanation! To seek a predictive explanation in the philosophy of mathematics is to misconstrue the discipline entirely. This thought might be extended to the conclusion that philosophical explanations lack any kind of knowledge (or unique belief) inducing property at all.[36]

4.5 Generalising the Picture

In this chapter we have tried to demonstrate the importance of being able to factor for structural information in cases of abductive reasoning. We began with a presentation of Gettier's argument to show that underdetermination problems can prompt an abductive problem which is used to undermine the notions of knowledge and non-question begging belief. We then proceeded to consider a case study in underdetermination in the philosophy of mathematics.

The idea was to show that if we considered the appropriate metaphysics then we could reexamine the underdetermination problem in the hope that we could resolve our abduction problem one way or the other. It became steadily clear that instead of resolving the problem we could at best dissolve the problem. This subtle distinction was motivated by examining the kind of dependency relationship which obtained between complex constructions in mathematics. On examination, it turned out that the grounding relationship often supposed to underlie metaphysical dependency of the relevant kind, was either (1) straightforwardly inapplicable to the debate or (2) ultimately inapplicable, because fundamental issues of mathematical ontology remain unresolved.

We suggested that these kinds of consideration should be expressible at the epistemic level of explanation. To do so we found the need to distinguish between types of explanation so that an explanatory relation in the philosophy of mathematics could be seen as domain-appropriate. It is too easy to consider explanation as a simple relation between premises and conclusions. The crucial observation here is that explanation is not a relation, it is a process! A process which is subtly different in each domain of reasoning.

An explanation is a process of moving from our shared structural information, and the observation of structural connection between event types, to the inauguration of default rules of expectation, or evidentiary relevance and the conclusion that certain claims justify (or explain)[37]

[36] A similar suggestion was made by Eric Schliesser on the NewApps blog on July 17th 2013 in a post entitled "What is the reach of philosophical argument"

[37] In lieu of developing a systematic "logic of explanation" we treat justification and explanation as complementary notions. Whatever claim justifies ψ may be used to explain ψ. So if we have j : ψ where j encodes a formula χ such that χ justifies ψ then χ is also an abductive solution to an abductive ψ-problem. Both are answers to the why-ψ

others just when they track a dependency relation we have seen hold in the world. These justifications prompt the formation of beliefs and or knowledge where appropriate. An explanation is successful if upon deployment within in a community, the community converges upon acceptance of the theories in which the appropriate default rules, or evidence function $\mathcal{E}$ feature. The process is not so much the idea of reasoning by appeal to structural information, but by doing so in such a way that others are compelled by your arguments - compelled because they share the same structural information, and see your reasoning to be adequate.

To illustrate the manner in which justification and indeed explanation is a process, we should note the manner in which justification terms or analogously explanatory terms can inherit properties from the dependency relation which they track. For instance, consider a language of explanatory grounding $\mathcal{L}_G$ from three tiers of linguistic utterance. For convenience we list them here: $\mathcal{L}_{pg}$, $\mathcal{L}_{exp2}$, $\mathcal{L}_{exp3}$. The idea is construct our logic of justification from bottom up, by defining justification with respect to our structural information as determined by, for instance, our logic of full ground.

$$\phi ::= \mathrm{p} \mid \neg\phi \mid \phi \vee \psi \mid \mathrm{j}\!:\ \psi$$
$$\phi ::= \mathrm{j} : \psi \mid \mathrm{j} + \mathrm{r} : \psi \mid \mathrm{j}! : \mathrm{j} : \psi \mid (\mathrm{j} \cdot \mathrm{e}) : \psi$$
$$\phi ::= \mathrm{p} \mid \phi \precsim \psi \mid \Delta \precsim \psi$$

Where the language $\mathcal{L}_G$ is constructed in three stages. We take that L_{pg} is primitive as it stems from our implicit logic of grounding. We, then define a meta-syntactic function in analogy with the CS-function of justification logic. The idea is provide an agent with a limited but well motivated set of justificatory claims from which we can motivate belief and knowledge claims. Each justification is based on a result achieved in our metaphysical theory of grounding.

$$(\phi_i) \in \mathcal{L}_G = \begin{cases} j : \psi \in \mathcal{L}_{exp3} & \text{if j} : \psi \in \mathcal{L}_{exp2} \\ Bool(\phi_i) \in \mathcal{L}_{exp3} & \text{if } \phi_i \text{ is a boolean combination of atomics} \\ p \in \mathcal{L}_{exp3} & \text{if } \phi_i \in \mathcal{L}_{pg} \text{ and is atomic} \\ j : \psi \in \mathcal{L}_{exp2} & \text{if } \vdash_{pg} \chi \precsim \psi \text{ and there is coding } j := \chi \\ j + r : \psi \in \mathcal{L}_{exp2} & \text{if } \vdash_{pg} \chi \cdot \tau \precsim \psi \text{ and there is a coding j } := \chi, r := \tau \\ j! : j : \psi \in \mathcal{L}_{exp2} & \text{if } \vdash_{pg} \tau \precsim \chi \wedge \chi \precsim \psi \text{ and there is coding j} := \chi,\ \text{!j } := \tau \\ j \cdot e : \psi \in \mathcal{L}_{exp2} & \text{if } \vdash_{pg} (\chi \precsim \tau \wedge \tau \precsim \psi)\text{where } j := (\chi \precsim \tau), e := \chi \\ \phi_i \in \mathcal{L}_{pg} & \text{if } \top \end{cases}$$

Crucially, we can see now that the properties of the justification relation e.g. monotonicity etc, are derived from from the underlying relation on which the justification procedure is founded. In our case, the non-monotonicity of the full grounding relation should ensure that the **monotonicity** axiom of justification logic will fail. This syntactic information can be encoded semantically in an evidence function $\mathcal{E}$ as appropriate. This picture of ascent from our base ontological and metaphysical theories to our reasoning about justification and belief is the crucial missing component in formal epistemology as standardly developed. The mistake oft repeated is to think that beliefs and knowledge can be studied in vacuum. The defence for the unreasonable idealisation is that it becomes too messy to incorporate such information. We would like to think that we have shown two things (1) the problems emerging from our failure to incorporate such structural information in our epistemology are much worse than any worry about theoretical nicety, and (2) by viewing justification (and analogously explanation) as a process of tracking rigorously definable dependency relations we need not be too concerned about the incursion of messiness.

questions.

In the above construction we have effectively stipulated evidentiary relevance by fiat, but there is another more subtle method which would achieve analogous results. The idea is that justification logic encodes justification terms based on the relation of logical entailment i.e. proof. The Lifting Lemma shows that we can construct justification terms which precisely encode the conditions under which the consequences are logically entailed by our premises. Hence, by analogy we might hope to prove a "Lifting Lemma" for each such dependency relation. In other words we might think that once we observe conditions under which our consequences are grounded in our premises, we could determine the type of information required to properly construct justification terms for reasoning about grounding. So there could be a construction which shows for each conclusion, how the conclusion followed from our premises by appeal only to the inference rules of Fine's logic of pure ground.

Some issues remain. We have been intent to demonstrate the importance of structural information in both the development of an ontological picture, but also as a constraint on the types of justificatory (or explanatory) procedure we develop. However, we have barely scratched the surface with respect to systematically investigating the interactions of structural information upon our doxastic and epistemic states. Perhaps worse, we have not fully developed a picture of how this works in a multi-agent epistemic setting. This will be crucial if we expect to see explanations as a process of justification which converges to community consensus. For instance, we might need to insist that explanatory solutions can only develop and converge in a community if they share the same basic structural information and assumptions. We do not hope to adequately address either of these important issues within the scope of this thesis. Both raise significant and non-trivial questions which we defer for later work.

In the next chapter we shall focus on examining the relation of causality in an effort to discern its structure and the type of role it has in both ontological and epistemic reasoning. We wish to underline the importance of structural information in rational inference. We also seek to further elaborate the notion of explanation as a process of reasoning from structural information to the appropriate changes of epistemic and doxastic states. As such the next chapter will provide a model that might be emulated in the development of further explanatory notions. Entirely in line with our current suggestions.

Chapter 5

Because: Identified Dependency

On the plain behind him are the wanderers in search of bones...and they move haltingly in the light like mechanisms whose movements are monitored with escapement and pallet so that they appear restrained by a prudence or reflectiveness which has no inner reality, and they cross in their progress one by one that track of holes that runs to the rim of the visible ground and which seems less the pursuit of some continuance than the verification of a principle, a validation of sequence and causality - as if each round and perfect hole owed its existence to the one before it there on that prairie upon which are the bones and the gatherers of bones...He strikes a fire in the hole and draws out his steel. Then they all move on again. - Cormac McCarthy[1]

5.1 Introduction: Patterns of pattern seeking

Seek and you shall find. This phrase, once a religious benediction, is now an infamous fallacy named optimism. Nevertheless, we shall argue, there are reasons to be optimistic.

In the previous chapter we argued that structural information had a role to play in epistemology. In particular we argued that certain dependency relations should be included in our explanatory reasoning. We attempted to showcase this by appeal to the relation of grounding. The choice was appropriate at the time, but we would be remiss if we did not mention what many[2] feel to be the paradigmatic explanatory relation - causality. We shall argue that the world has a discernible causal structure, that our best theories are those which seek and find such information.

We do not pretend to address every available theory of causality. We deliberately narrow our focus to avoid over-inflating the need for exposition. In particular we treat Judea Pearl's *Causality*[3] in addition with two papers he wrote with Joseph Halpern on the same topic. Our intent is to view the study of causality as a prelude the study of explanation.[4] The primary reason for making this choice is that Pearl and Halpern take seriously the notion of structural information and they are rigorous about distinguishing causal structural information from statistical data. These reasons are sufficient, but we provide another. Pearl explicitly links the recognition of true causal information with the emergence of a "deep understanding."[5] He relates the observation of structural information with the growth of understanding and the emergence of knowledge and belief. Since this is exactly the course we recommend, we shall proceed to assess his proposal.

We shall break down the discussion into three areas where the work of Pearl and Halpern

[1] In *Blood Meridian* the Epilogue.

[2] cf. [66]

[3] [73]

[4] In this choice we follow Pearl and Halpern in [47]

[5] For the smoothest introduction to Pearl's thinking on this issue see the Epilogue in [73]

can be seen to be innovative. In particular we shall examine (1) their discussion of the graphical representations of dependency information (2) graphical conditions for the identification of causal relationships (3) their semantics for causal statements. All three contribute towards the broad theory of causality and (1) and (2) are utilised to develop a theory of explanation derivative of the predictive power of causal reasoning.

5.1.1 Mathematical Preliminaries

In this chapter we wish to represent structural information, in particular structural dependence and independence. We begin with the idea that relations of probabilistic independence can be represented in the form of graph. This is convenient, but furthermore it allows us how to switch casually between probabilistic estimates of dependence and qualitative representations of these relations thereby cementing the idea that reasoning about uncertainty should be able to factor for both qualitative and quantitative information.

Independence

We define a relation of independence. Any pair of events which is not independent is taken to be dependent to some degree. The challenge, as always, is learning to specify what degree of dependence holds amongst observed events.

Definition (Independence with respect a CPS) H and T are probabilistically independent with respect to a conditional probability space (Ω $\wp(\Omega)$, Ω' σ) if H $\in$ Ω' implies that σ(T | H) = σ(T) and T $\in$ Ω' implies σ(H | T) = σ(H).

We often see what is known as the multiplicative definition of independence which states that σ(H $\cap$ T) = σ(H)σ(T). Fortunately we can show this to be equivalent to the above conditional definition. Suppose that σ(H $\cap$ T) = σ(H)σ(T) and that σ(T) $\neq$ 0, just to avoid the trivial case. We want to show that σ(H| T) = σ(H), but σ(H | T) = $\dfrac{\sigma(H \cap T)}{\sigma(T)}$, which by assumption is the same as $\dfrac{\sigma(H)\sigma(T)}{\sigma(T)}$. This cancels giving us σ(H). By the transitivity of identity we are done. For the other direction suppose that σ(H) = σ(H | T). We want to show that σ(H$\cap$T) = σ(H)σ(T). Assume σ(H$\cap$T), then by the multiplication rule[6] we have σ(H |T)σ(T). By our initial assumption this is the same as σ(H)σ(T), as desired. From this observation we can prove the following theorem.

Theorem If T and H are independent then so are T^c, H and T, H^c and T^c, H^c.
We work with the multiplicative definition of independence. Observe: $\sigma(H^c \cap T) = \sigma(T \setminus H) = \sigma(T) - \sigma(H \cap T) = \sigma(T) - \sigma(H)\sigma(T) = (1 - \sigma(H))\sigma(T) = \sigma(H^c)\sigma(T)$. Given this result the other equivalences follow by a similar proof.

Theorem (The Chain Rule) σ((H$_1$)$\cap$..... $\cap$(H$_n$)) = σ(H$_1$)σ(H$_2$ | H$_1$)σ(H$_3$ | H$_1$$\cap$H$_2$)....$\sigma$(H$_n$ | H$_1$$\cap$.....$\cap$(H$_{n-1}$).
The proof follows by observing that:

$$\sigma(\mathrm{H}_1)\sigma(\mathrm{H}_2 \mid \mathrm{H}_1)\sigma(\mathrm{H}_3 \mid \mathrm{H}_1\cap\mathrm{H}_2)....\sigma(\mathrm{H}_n \mid \mathrm{H}_1\cap.....\cap(\mathrm{H}_{n-1}) = \sigma(\mathrm{H}_1)\frac{\sigma(H_2 \cap H_1)}{\sigma(H_1)}$$
$$\frac{\sigma(H_3 \cap H_2 \cap H_1)}{\sigma(H_1 \cap H_2)}\cdots\cdots \frac{\sigma(H_n \cap ... \cap (H_1)}{\sigma(H_1) \cap .... \cap (H_{n-1})}$$

Multiplying out we see that for any conditional claim in the sequence the numerator cancels with the denominator of the previous conditional in the sequence. Hence, we are left with just the

[6] Discussed in chapter three.

numerator of the nth conditional in our sequence. In other words, we have shown the equivalence $\sigma((\mathrm{H}_1)\cap\ldots\ldots \cap(\mathrm{H}_n)) = \sigma(\mathrm{H}_1)\sigma(\mathrm{H}_2 \mid \mathrm{H}_1)\sigma(\mathrm{H}_3 \mid \mathrm{H}_1\cap\mathrm{H}_2)\ldots.\sigma(\mathrm{H}_n \mid \mathrm{H}_1\cap\ldots\ldots\cap(\mathrm{H}_{n-1})$ as desired.

Definition (Conditional Independence with respect to a parameter) H is probabilistically independent of T when conditional on E with respect to σ: written I_σ(H, T | E) if $\sigma(\mathrm{T} \cap \mathrm{E}) \neq 0$ implies $\sigma(\mathrm{H} \mid \mathrm{T} \cap \mathrm{E}) = \sigma(\ \mathrm{H} \mid \mathrm{E})$. In words, learning T is irrelevant to determining the value for H given E.

The natural analogue of the multiplicative definition of independence in this setting the following claim: H is independent of T given E iff $\sigma(\mathrm{H} \cap \mathrm{T} \mid \mathrm{E}) = \sigma(\mathrm{H} \mid \mathrm{E})\sigma(\mathrm{T} \mid \mathrm{E})$. Again this is provably equivalent to the above definition of conditional independence. This defintion is often convenient for proving the following important properties:

Symmetry If I_σ(H, T |E) then I_σ(T, H | E)

Contraction if I_σ(H, T | E) and I_σ(H, S | T ∪ E) then I_σ(H, T ∪ S | E).

Weak Union if I_σ(H, T ∪ S | E) then I_σ(H, T | S ∪ E)

Decomposition if I_σ(H, T∪ S | E) then I_σ(H, T | E)

Composition if I_σ(H, T | E) ∧ I_σ(H, S | E) then I_σ(H, S∪T | E).

These properties are largely uncontroversial mathematically. However Pearl explicitly cites them as part of a kind of axiom set for reasoning about dependency relations on graphs.[7] We will show below how these properties can be represented graphically, but first consider what they mean. *Symmetry* states that if H is independent of T given E, then so too is T independent of H given E. Pearl paraphrases *Weak Union* in terms of relevance stating that the axiom ensures that "learning irrelevant information S cannot help irrelevant information T become any more relevant to H." The paraphrase is suggestive in that we might now think to read the *Symmetry axiom* as stating that if T is irrelevant to H, then H is irrelevant to T. Or analogously, it we learn nothing about H from T we can learn nothing about T from H. The *Contraction* axiom states that if we learn that some S is irrelevant to H, after learning some T, then S must have been irrelevant to H before we learned T also. Whereas *Decomposition* states that if S and T are together irrelevant for the truth of H, then they are also irrelevant separately.[8] Pearl calls these the *graphoid* axioms.

We quickly show how to prove that Symmetry is valid on any CPS. Due to the nature of the independence predicate I_σ we need only prove an equivalence result. Assume $\sigma(\mathrm{H} \mid \mathrm{T} \cap \mathrm{E}) = \sigma($ H | E). We need to show that $\sigma(\mathrm{T} \mid \mathrm{H} \cap \mathrm{E}) = \sigma(\ \mathrm{T} \mid \mathrm{E})$.

(1) . $\sigma(\mathrm{T} \mid \mathrm{H} \cap \mathrm{E}) = \dfrac{\sigma(H \cap E \mid T)\sigma(T)}{\sigma(H \cap E)}$ By Bayes' Rule.

(2) $= \dfrac{\sigma(T \cap E \cap H)}{\sigma(H \cap E)}$ By the Multiplication rule and Rearranging.

(3) $= \dfrac{\sigma(E \mid T)\sigma(H \mid E \cap T)\sigma(T)}{\sigma(H \mid E)\sigma(E)}$ Rearranging and the Chain rule.

(4) $= \dfrac{\sigma(E \mid T)\sigma(H \mid E)\sigma(T)}{\sigma(H \mid E)\sigma(E)}$ By our Assumption.

[7] In [73]pg11-12. It was briefly thought that Symmetry, Contraction, Weak Union and Decomposition might be sufficient to achieve a finite characterisation of independence relations. This dream was overturned by Studeny in [65], who also showed that there could be no finite characterisation of the Independence relation.

[8] The *Composition* axiom actually follows from the others. We may quickly prove it by assuming the truth of the antecedent and the falsity of the consequent, and then apply the contraposed *Contraction* axiom to derive a contradiction.

(5) $= \dfrac{\sigma(E \mid T)\sigma(T)}{\sigma(E)}$ By Cancelling

(6) $= \sigma(\mathrm{T} \mid \mathrm{E})$ by Bayes' Rule.

So again, by the transitivity of identity *Symmetry* holds on any conditional probability space. Similar results can be found for all the other properties listed above.[9] This is significant because it suggests the idea that we might be able to develop qualitative rules for reasoning about dependency relationships! At no point have we relied on particular numeric values. We will develop the idea that for any graph representative of a conditional probability measure, we know that the graphs must be closed with respect to the *graphoid* axioms. However you might worry that the difficulty which emerges is the fact that these rules rely on the notion that σ(X) is unfixed. Most problems, you claim, emerge because we cannot accurately predict how σ(X) becomes fixed. This worry is needless. Uncertainty itself is not a vice, what is important is that we can derive qualitative rules of inference even when the parameters are unfixed. In doing so we can reason well about species of non-logical dependency by adding or removing certain axioms as our model requires.

Random Variables

This latter point is nicely exhibited by Halpern who discusses the type of conditional probability space wherein each subset of the space can admit multiple realisations.[10] As we shall show, this feature does nothing to negate the idea that we can successfully draw up qualitative rules to govern expectation.

Definition (Independence with Random Variables) Let Val(X) be the set of possible realisations of the event/propositional variable X. We say that variables X and Y are *(probabilistically) conditionally independent* given Z with respect to the probability measure σ if, for x $\in$ Val(X), y $\in$ Val(Y) and z $\in$ Val(Z), the event X = x is conditionally independent of Y = y given Z = z. We say a Random variable is discrete if it only has finite potentially realised states.

Definition (Conditional Independence with sets of random variables) Let $\mathbf{X} = \{ \mathrm{X}_1 .... \mathrm{X}_n \}$ $\mathbf{Y} = \{ \mathrm{Y}_1 .... \mathrm{Y}_m \}$ and $\mathbf{Z} = \{ \mathrm{Z}_1 ... \ \mathrm{Z}_k \}$. Then we say that **X** and **Y** are conditionally independent given **Z**, with respect to σ, written $I^{rv}_{\sigma}(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z})$ if $\mathrm{X}_1 = \mathrm{x}_1 \cap .... \cap \mathrm{X}_n = \mathrm{x}_n$ is conditionally independent of $\mathrm{Y}_1 = \mathrm{y}_1 \cap ... \cap \mathrm{Y}_m = \mathrm{y}_m$ given $\mathrm{Z}_1 = \mathrm{z}_1 \cap ... \cap \mathrm{Z}_k = \mathrm{z}_k$ for all respective elements of the appropriate indices. This definition works because there it is possible to find an intersection to collection of overlapping events, or similarly a shared truth amongst collections of beliefs/propositions.

As an example, Halpern argues that the natural analogue of *Weak Union* is valid on any conditional probability space able to account for random variables. In other words: If $I^{rv}_{\sigma}(\mathbf{X}, \mathbf{Y} \cup \mathbf{Y'} \mid \mathbf{Z})$, then $I^{rv}_{\sigma}(\mathbf{X}, \mathbf{Y} \mid \mathbf{Y'} \cup \mathbf{Z})$. Similar remarks apply to the other proposed axioms. The proofs are not too difficult but they lie beyond the scope of this chapter. We now to turn to how to best represent such information.

5.2 Bayesian Networks

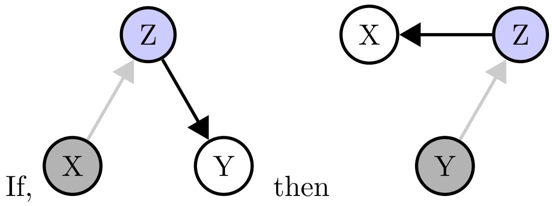
The idea is that we can represent structural information on graphs in which the arrows between the nodes represent, or as can happen, fail to represent a relation of dependence amongst the nodes - where each node is thought of as a event.

[9] A proof that all these axioms are indeed valid in a conditional probability space is contained in Wolfgang Spohn's paper [85] section 5. Most of the axioms are no more difficult than the above proof of *Symmetry.*

[10] See [45]pg129 -132.

Qualitative Bayesian Networks

Consider this picture.



We have been arguing that the notion of conditional independence can be well defined, albeit not exhaustively defined with finite specifications. We are now in position to show this information may be represented graphically. Allow that the blue nodes represent the information we are conditionalising upon, and the grey nodes represent the now irrelevant information. We also let the black arrow represent direct influence. The picture on the left reads as follows: Y is independent of X given Z. So by the *Symmetry* axiom, we are entitled to conclude that X is also independent of Y, given Z. This is exactly what our second picture indicates. Lets make this more precise.

Definition (Graph) A graph G is a pair (G, $\to$) in which G is a set of nodes and $\to$ is a binary relation over G. The relation between nodes is often called an edge, an edge is directed if there is an arrow pointing from the parental node to the child node for every such pair. We say that is Y $\in$ G *is a child of* X $\in$ G iff X$\to$Y. Similarly, we say that X $\in$ G *is a parent of* Y $\in$ G iff X $\to$ Y.

We think of a graph G as representing a qualitative ordering relation on a state space Ω by treating Ω as a set of possible worlds containing n binary random variables $\mathcal{X}$ = { X$_1$....X$_n$}. So that w $\in$ Ω is a tuple (X$_1$ =x$_1$....X$_n$ = x$_n$) where x$_i$ $\in \{0, 1\}$ = Val(X$_i$). This is tantamount to a representation of dependencies amongst propositional information. Each graph can be thought of as representing a (typically partial) world-history. Since we wish to use graphs to represent relations of causal influence, we should put certain constraints on the graph. Most notably, we deny that any fact in the history of the world can be its own cause, so there should be no circularity in our causal diagrams.

Definition (DAGs) A directed acyclic graph is a graph with directed edges such that there does not exist a sequence of nodes n$_1$.....n$_k$ such that n$_1$= n$_k$ and there is an edge from n$_i$ to n$_{i+1}$ for i = 1,...., k-1.

So far so good, but how does this help us? We shall show that we can use these DAG structures to motivate our beliefs in conditional probabilities. The idea is that we begin with qualitative maps of dependency structures as represented on a DAG, the observation of which can be used to induce quantitative beliefs about the conditional dependencies among certain states of affairs.

Trading on the intuitive notion of an ancestor relation we can define two more parameters for any DAG. Observe that your parents, had their parents and so on...., similarly with some generous assumptions you might hope that your children will have their own children and so on... Note that you are directly causally responsible for your own children, but you are not the cause of your ancestors or the direct cause of you ultimate descendants. This would suggest that we are working with a notion of intransitive causality. But we need not commit to that just yet.

Definition (Descendants and Non-Descendants) Given a Qualitative Bayesian network G, let the Par$_G$(X) be parents of the random variable X $\in$ G. Let Des$_G$(X) be the descendants of X, i.e. the set of all those nodes Y such that X is an ancestor of Y. It's now easy to see that we can define the set of Non-descendants of X as NonDes$_G$(X) = $\mathcal{X}\backslash$Des$_G$(X).

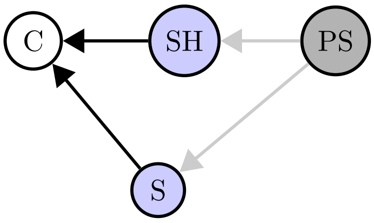
With these distinctions in mind we can finally draw the connection between qualitative and quantitative representations of dependence and independence.

Definition (Qualitative Representation) The Bayesian network G is compatible with the probability measure $\sigma \in \mathcal{P}$ if $I_\sigma($ X, NonDes$_G$ | Par(X)). This latter condition is called the Causal Markov condition.[11]

So clearly a graph shows X to be conditionally independent of its non-descendants given its parents for all X $\in \mathcal{X}$. Note that all our indirect ancestors are amongst the set of our non-descendants. You can see that this holds in the above pictures, and that therefore we can find a probability measure to assign to each graph. Perhaps more importantly, we can now see that each parental-child relationship as a deterministic function which qualitatively determines the state of the child given the state of the parents.[12] Rips criticises the Markov condition on the basis that there is "little positive evidence" that people reason causally in about an intransitive causal relation.[13]. We will consider this objection later when we elaborate Pearl's notion of counterfactual reasoning.

5.2.1 Quantitative Bayesian Networks

To see that our graphical structure can be used to represent further quantitative information we need only precisely quantify the degree of influence a parental node enacts on its children.[14] Consider the following picture:



The idea here is that we have a qualitative map of the influence between smoking and cancer. Where C is the event of contracting cancer, S is the habit of smoking, and SH is the occurrence of being exposed to second hand smoke. The claim above is that contracting cancer is independent of whether or not your parents smoked PS, given the fact that you both smoke S and are often exposed to second hand smoke SH.

We move from a qualitative statement to a quantitative one by ascribing a function f to our qualitative Bayesian network, which generates a table that quantifies the influence of every parent-child relationship for all nodes in G. So formally, a quantitative Bayesian network is (G, f), where G is a qualitative Bayesian network and f is a function $f\colon$ Val(Parents) $\mapsto \sigma($Child = c$_i \in$ Val(Child)) where $\sigma \in \mathcal{P}$. Obviously, we can provide real figures to substantiate the independence claim above,

[11] Glymour *et al* note in [40]pg9 that "the Markov condition is not given by God" but it is often quite reasonable. We agree with this assessment. The assumption is weak enough as to be plausible in many discussions of causal dependence. However, Burnett and Medin note that there is some evidence that scientists working in various fields often override the assumption in [1]pg950-951

[12] For a more detailed discussion see [73]pg30-32

[13] He ultimately suggests that we need a less "nerdy" representation of causal reasoning less tied to statistical dependencies and more attentive to discursive (or semantic) evidence. See [1]pg624-625

[14] Traditionally, we would draw up a conditional probability table assigning the degree of influence parents have on their children. So for example if the parents of X in G are Y and Z, the cpt for would have an entry spelling out the realisations of Y and Z respectively, so that we determine σ(X = 1 | Y = y $\cap$ Z = z) on the basis of those realisations. So perhaps our parents are realised in such a way, that for them to ensure X = 1, they have to act in concert i.e. they must themselves realise the same value either 1 or 0, and if only one parental node instantiates 1, then for instance σ(X = 1) = .5. The situation becomes more complicated depending on the number of potential realisations our parental nodes can come to instantiate.

but note that the claim is supported by the quantitative data, and not derived from it. This is not to say that we never extract causal information from statistical data, but merely that causal claims can be understood in their own right, and indeed should be.

Definition (Quantitative Representation) A quantitative Bayesian network (G, f) represents the probability measure $\sigma \in \mathcal{P}$ if G qualitatively represents σ and the conditional probability table determined by f agrees with σ in so far as for each random variable X, the allocation of a probability for X conditional on the values for it's parents determined by f is the same as would be determined by σ.

Given our set up it should be easy to see that we can move back and forth between conditional probability measures and quantitative Bayesian networks. Take any conditional probability measure and once we have determined the independence relationships amongst our events, we can reconstruct this as a quantitative Bayesian network.[15] This result was first proved by Verma and Pearl in their paper "Causal Networks". The proof is almost immediate from the definitions.[16] The more significant result of the soundness and completeness of the graphoid axioms with respect to the structures of CPS was achieved by Geiger *et al* in their paper "Identifying Independence in Bayesian Networks"[17] We need to define one other notion before this result becomes tractable.

5.2.2 The d-separation Criterion

So far we have been content to translate our graphs into the vocabulary of probability to determine whether or not certain dependency relations hold, but it is possible to read off this information directly from the structure of our graph. The idea is that we wish to say when X is conditionally independent of Y given **Z** by examining only the structure of the graph.

Definition (d-separation) A node X is *d-seperated* (or blocked) from a node Y by a set of nodes **Z** in G, written $d\text{-}sep_G$(X, Y | **Z**) if for every undirected path from X to Y , there is a node Z' $\in$ **Z** such that :

1. We have on the path (a) X → Z' → Y or a fork (b) X ← Z' → Y, where an element of z interrupts the path from x to y and vice versa, or

2. We have on the path X → W ← Y, and neither W or any of its descendants are in **Z**

An undirected path is one which ignores the direction of the arrows. Generalising this definition we can say **X** is d-separated from **Y** by **Z** if every X in **X** is d-separated from every node Y in **Y** by **Z**. The first clause is straightforward, in (a) the element Z' blocks the influence of X on Y, and vice versa by symmetry, while in (b) it's clear from the fact that since we have no causal influence on our parents we cannot expect to be the cause of our parents' other children. The second clause is a bit less obvious. Why should we think that X is independent of Y because they both cause W? Recall our example about the causes of cancer. Since both S and SH are represented as being sufficient causes of cancer the presence of one should lessen the probability of the other, and since neither C or any of descendants (i.e. Death) are sufficient causes of either our smoking habits, or the presence of second hand smoke, we can be sure that S is independent of SH. Evidently, this

[15] There are a number of ways to proceed with such a construction. Halpern discusses the notion of a well-designed Bayesian network as one which recovers the dependence relations in an appropriate order so that the causal dependencies are explicitly tied up with the parent child relation.The broad idea is to take all the independence information determined by σ by noting the relevant formulas of the form I_σ(X, Y | Z), and order the dependencies such that when mapping events into nodes on a graph no child proceeds their parent in the graph. For details see [45] or Theorem 3.3 in [40]

[16] See Theorem 1 in [95]

[17] [37]

definition is in exact congruence with the idea that only parental nodes causally effect their children.

With this notion of independence in mind, we may think of our Bayesian Networks as being closed under the graphoid axioms as a proof theory, and conditional probability spaces as our model theory for reasoning about causal dependence. We've already indicated that the so called graphoid axioms are validated on a basic conditional probability space, but they were properly speaking formulated to define the notion of independence resulting from the d-separation criterion. Indeed we can see that the d-separation criterion allows us to prove other properties independence. For instance: we can also prove that a kind of *Weak transitivity* holds where G $\vdash$ *d-sep*$_G$($\mathbf{X}$, $\mathbf{Y}$ | $\mathbf{Z}$) $\wedge$ *d-sep*$_G$($\mathbf{X}$, $\mathbf{Y}$ | $\mathbf{Z}$∪Z) $\Rightarrow$ *d-sep*$_G$($\mathbf{X}$, Z | $\mathbf{Z}$) $\vee$ *d-sep*$_G$(Z , $\mathbf{Y}$ | $\mathbf{Z}$).[18] The notion that we can prove properties of certain independence relations graphically prompts a question; what is the relation between graphical proofs in DAG structures and probabilistic proofs in CPS structures? The next result answers that question.

Theorem (Soundness and Completeness with respect to CPS)

Soundness If *d-sep*$_G$($\mathbf{X}$, $\mathbf{Y}$ | $\mathbf{Z}$) then I_σ($\mathbf{X}$, $\mathbf{Y}$ | $\mathbf{Z}$) where σ is compatible with G.

Completeness If I_σ($\mathbf{X}$, $\mathbf{Y}$ | $\mathbf{Z}$), holds for $\sigma \in \mathcal{P}$ compatible with G, then *d-sep*$_G$($\mathbf{X}$, $\mathbf{Y}$ | $\mathbf{Z}$).

Soundness: Proof Sketch We rely heavily on the compatability definitions. Suppose that G $\vdash$ *d-sep*$_G$($\mathbf{X}$, $\mathbf{Y}$ | $\mathbf{Z}$). Then we need to show that CPS $\models$ I_σ($\mathbf{X}$, $\mathbf{Y}$ |$\mathbf{Z}$) for a compatible σ. It is always possible to find a compatible σ by the observations above. The proof of soundness is by induction on the depth of the graph G. For a graph G with one node X, then trivially, our CPS is σ compatible with the single event X and any independence claims validated in our G are provable in our CPS trivially.

For our induction hypothesis we suppose that for all G-structures with less than k-nodes, we can find a G-compatible σ function.

For the k-case. We look at the G$_k$ structure and proceed to construct the "well-designed" corresponding model CPS$_k$ so as to maintain the parental-child relation. Pick the "youngest descendent" in our chain, call it "Kid" and remove it, from CPS$_k$, i.e. let CPS$_{k-1}$ = CPS$_k$-[Kid] so that all the triplets $I\sigma$(X, Y | Z) involving [Kid] are not modelled in CPS$_{k-1}$. But then by our Induction Hypothesis CPS$_{k-1}$ is compatible with the graph G$_{k-1}$.

Now, note that for any triplet modelled in G$_k$, If G$_k$ $\vdash$ *d-sep*$_G$($\mathbf{X}$, $\mathbf{Y}$ | $\mathbf{Z}$), then there are three cases: either (i) [Kid] does not occur in the formula (ii) [Kid] occurs in the first or second entry (iii) [Kid] occurs in the third entry. We must show for each case CPS$_k$ $\models$ I_σ($\mathbf{X}$, $\mathbf{Y}$ | $\mathbf{Z}$), and then conclude that CPS $\models$ I_σ($\mathbf{X}$, $\mathbf{Y}$ | $\mathbf{Z}$). Thereby showing that at every stage of the construction of G, the validities of G are preserved, thereby by proving compatibility of our construction CPS with G. This will conclude our proof of soundness.

We sketch the ideas underlying two cases[19] : (i) [Kid] does not occur in the formula. Hence the formula is modelled in G$_{k-1}$ so by our induction hypothesis CPS$_{k-1}$ $\models$ I_σ($\mathbf{X}$, $\mathbf{Y}$ | $\mathbf{Z}$) with a G$_{k-1}$ compatible function σ. Note that CPS$_{k-1}$ $\subset$ CPS. So by the nature of our mapping construction CPS $\models$ I_σ($\mathbf{X}$, $\mathbf{Y}$ | $\mathbf{Z}$) as desired. The idea behind the proof of (iii) is as follows: Observe that if we d-seperate $\mathbf{X}$ and $\mathbf{Y}$ with respect to [Kid] ∪ $\mathbf{Z}$ in G$_k$, where [Kid] is a sink i.e. has no outgoing arrows due it's relative youth, then we must also d-seperate $\mathbf{X}$ and $\mathbf{Y}$ with respect to $\mathbf{Z}$

[18] This result is proven by Dan Geiger in [36]. Intuitively, the weak transitive axiom states that if two subsets of the statespace are both conditionally and unconditionally independent of each other give a singleton variable Z, then it is impossible for both $\mathbf{X}$ and $\mathbf{Y}$ to be dependent on Z. It's important to note that the CPS has to be augmented with some basic algebraic properties to validate the most general form of the *graphoid* axioms with sets of random variables. For details see discussion in [45]pg 131-132.

[19] We omit case (ii) largely because of need for brevity.

by *Decomposition.* So $\text{G}_k \vdash$ *d-sep*$_G(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z}) \wedge$*d-sep*$_G(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z}\cup[\text{Kid}])$. By the definition of d-seperation and the consequent *Weak-transitivity* this result ensures that $\text{G}_k \vdash$ *d-sep*$_G(\mathbf{X}, [\text{Kid}] \mid \mathbf{Z}) \vee$ *d-sep*$_G([\text{Kid}], \mathbf{Y} \mid \mathbf{Z})$. Hence, we also have that $\text{G}_k \vdash$ *d-sep*$_G(\mathbf{X}\cup[\text{Kid}], \mathbf{Y} \mid \mathbf{Z}) \vee$ *d-sep*$_G(\mathbf{X}, \mathbf{Y}\cup[\text{Kid}] \mid \mathbf{Z})$ by *Symmetry* and *Composition.* Putting all this together, we can show (by our proof of the second case (ii)) that $\text{CPS}_k \models I_\sigma(\mathbf{X}\cup[\text{Kid}], \mathbf{Y} \mid\mathbf{Z}) \vee I_\sigma(\mathbf{X}, \mathbf{Y}\cup[\text{Kid}] \mid \mathbf{Z})$. In either case we can apply *Weak Union* and *Symmetry* accordingly to swap [Kid] back into the third entry, so as to conditionalise on [Kid] as required.

In the discussion above regarding Soundness, we deliberately hedged on the details about constructing an appropriate CPS for any given graph G. However the crucial feature of the proof for completeness is the manner in which we construct the model. Where our hedging in the case of soundness is excusable because the construction is relatively straightforward, the details of the completeness proof are better left unstated rather than quickly stated.[20]

5.3 Causality and Model Testing

So far we have argued that we can switch between Qualitative and Quantitative reasoning about causal dependence. We now present an argument to the effect that the qualitative reasoning is more basic, or precedes the quantitative reasoning. We then demonstrate how we might develop a theory of causation on the ability to test models for accuracy.

5.3.1 The Suppression Test

Ruth Byrne proposed and conducted a study in the psychology of reasoning.[21] Participants are presented with a reasoning task of the following form: They are given two premises (1) If Jane has an essay to write she will study late in the library. (2) Jane has an essay to write. In 90 percent of the cases participants followed the standard Modus ponens inference pattern and concluded that (C) she will study late in the library.

At the next stage of the experiment, we introduce a third premise (3) If the library is open, she will study late in the library. At this point only 60 percent of the participants, conclude that she will study late. This phenomenon is called the suppression effect for modus ponens. Byrne seeks to use this experiment to undermine the notion that purely classical logic is an empirically accurate truth-tracking mechanism. Stenning and van Lambalgen suggest instead that the moral to be drawn is that the nature of our formal inference is not fixed prior to interpretation, but rather imposed after the interpretation given our understanding of the content of the premises involved.[22]. In particular the idea is that upon receipt of the third premise (3) we decide to interpret the argument in a default logic.

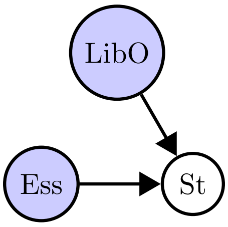
If we try to formulate the reasoning in terms of probabilities, we cannot predict the suppression effect unless we directly factor for the causal impact. In other words we must insist that the closure of the library ensures that Jane cannot go to the library regardless of whether she has an essay. As such, we need to factor more directly for the causal effect of particular event types - that is to

[20] Again, full details can be found in Geiger's Phd thesis [36]. We provide an outline: *Completeness: Proof Outline* The proof is by contraposition. We suppose that it's not the case that *d-sep*$_G(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z})$. Then we aim to show that there is a G-compatible function σ where the induced CPS is such that CPS $\not\models I_\sigma(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z})$. By assumption there is no d-separation between $\mathbf{X}$ and $\mathbf{Y}$, so there is an active path, between the two. For each interaction on that path, we have to correlate the parental-child relations with a compatible distribution σ and ensure that causal influences flows along that path whatever the realisations of the (potentially) intervening nodes. Preserving these features requires a complex construction, we shall not elaborate.

[21] See discussion in [90]pg180

[22] See [90]

say, we should factor directly for causal dependencies if we hope to render probabilistic reasoning a faithful model of empirically observable inference patterns.[23] Consider the Bayesian network which states that Jane's ability to study is dependent on her having an essay and the library being open.



Only once we have assigned a compatible probability function to the network can we expect to predict the appropriate suppression. In particular we want to insist on the fact that $\sigma($ St $\mid$ Ess $= 1 \wedge$ LibO $= 0) = 0$, so that even if $\sigma($St $\mid$ Ess $= 1 \wedge$ LibO $= 1)$ is high, we can still expect to see some suppression effects, unless our participants explicitly believe $\sigma($LibO $= 1)$ is also high. Compliance with our underlying causal intuitions is the criteria of adequacy for finding a compatible σ-function. Pearl argues similarly that "the original authors of causal thought cannot be ignored when the meaning of the concept is in question. Indeed, compliance with human intuition has been the ultimate criterion of adequacy in every philosophical study of causation"[24]

5.3.2 Simpson's Paradox

In addition to underlying our probabilistic assignments our causal information can be seen to solve some standardly arising riddles in statistics.

Allow that we have 26 participants in an experiment - 13 female, and 13 male. We have two drugs D_1 and D_1. Eight women use the drug D_1 and two recover, and five men use D_1 while one recovers. At the same time five women and eight men use drug D_2 while four women and six men recover. For drama let us assume that everyone untreated (i.e. not participants) dies. So we can recover the two following results:

1. $\sigma(R \mid F \cap D_1) = \dfrac{\sigma(R \cap F \cap D_1}{\sigma(F \cap D_1)} = \dfrac{2}{8} = 0.25$
2. $\sigma(R \mid M \cap D_1) = \dfrac{\sigma(R \cap M \cap D_1)}{\sigma(M \cap D_1)} = \dfrac{1}{5} = 0.2$

Hence, an inequality holds and the probability that a woman recovers when given the drug D_1 is greater than the probability that males recover. Similarly, we can compute the likelihood that a person recovers given their consumption of the second drug. From this computation we can see that the probability of a woman recovering from consumption of the second drug D_2 is greater than the probability of being a recovering male. With this result we should be happy to conclude that there is a greater probability of recovering if you are a woman. However, we can see that:

3. $\sigma(R \mid F) = \dfrac{\sigma(R \cap F)}{\sigma(F)} = \dfrac{2+4}{13} = \dfrac{6}{13} < .5$
4. $\sigma(R \mid M) = \dfrac{\sigma(R \cap M)}{\sigma(M)} = \dfrac{1+6}{13} = \dfrac{7}{13} > .5$

So it is apparently preferable to be male rather than female, but on inspection of the above inequalities we can see that it is preferable in each case to be a treated woman. Can we explain the apparent inconsistency? If we distinguish between evidential information and directly causal

[23] The argument is discussed with more detail in [90]pg215-216,

[24] See [73]pg26-27.

information we can make sense of these observations because Simpson's paradox shows that causation and statistical correlation come apart. Pearl has a similar discussion of Simpson's paradox in which he argues that we must distinguish between observed correlation, and net causal effect.[25] In short, we should be able to distinguish between the merely factual observation that σ(R | F $\cap$ D_1) and the law σ(R | F $\cap$ $do(D_1 = 1)$) = high, where the formula states that the probability that a woman recovers when she does take the drug D_1. The distinction stems from the intuition that there is an expectation of some causal effect, given the fixing of certain parameters. Such thinking allows us to understand why the inequality between (3) and (4) is less compelling than the inequality between (1) and (2), and how the inconsistency should be resolved in favour of our causal information.[26] In short, whether man or woman we should take the drugs, and refrain (where "needed") from a sex change operation.

5.3.3 Identification of Causal Effects by Model Preference Algorithms

In our first chapter we discussed the Grue paradox and proceeded to argue on the basis of Occam's razor that there were significant reasons to reject the grue-hypothesis. We can recover such an argument in this setting to develop a preference for simpler causal models if we can define the complexity of a causal model. We consider briefly an attempts to inaugurate this standard of model preference.

A Theory of Inferred Causation

The underlying idea of this section is that no amount of statistical information regarding the correlation between two events, will be sufficient to imply a causal connection between those two events. This has, since Hume, been the cross borne by any causal theorist - no amount of conformity with a proposed causal law is apt to ensure future conformity in all cases. As such we need to provide other criterion which can be seen to motivate the establishment of causal laws, otherwise all causal theories beg the question against a consistent but arbitrary alternatives. Paying heed to Hume's observation we shall define a causal model to include hidden variables which Nature is at liberty to have influence the observed variables. The idea is to "view the task of causal discovery as an induction game that scientists play against Nature. Nature possesses stable causal mechanisms that, on a detailed level of descriptions are deterministic functional relationships between variables, some of which are unobservable.These mechanisms are organised in the form of an acyclic structure, which the scientist attempts to identify from the available observations."[27] Importantly, the scientist imposes certain criteria of adequacy so as to derive the most preferable model of these mechanisms.[28]

Models of causal processes can have different degrees of granularity, where there exists a perfect causal model of the world which factors for all pertinent parameters we expect this model to meet the Causal Markov condition for all nodes. As we get less accurate models we abstract away from some of the details including "hidden variables" as explanatory devices under which the state of some node is a function from the hidden variable and its parents. These can be more or less harmful. They are harmless if we treat them as short hand for a process we could properly speaking model with appropriate time and effort, they are harmful if taken to be a mysterious, inexplicable and forever intangible feature of Nature. Pearl treats these hidden variables as latent albeit "unobserved" variables, the details of which we could flesh out if we so chose.

[25] See his account in [73]pg174 -180.

[26] Similar remarks apply directly to other covariate selection problems e.g. the grue-hypothesis on the basis of a detailed theory about our colour perception. For details see [73]pg128 -130.

[27] [73]pg43

[28] A more precise formal presentation of science as a game played towards a limit can be found in Kelly's [53]pg130-136

Definition (A Causal Structure) A DAG is a Causal structure where each link between the variable nodes represents a direct functional relationship among the relevant variables. Our causal structure is a set of nodes V which contains both observed and unobserved variables. The structure is closed under the → relation as expected.

Definition (A Causal model) M = (G, $Feat_G$) where G is a graph of a particular DAG, and $Feat_G$ are the features of the causal model G. So $Feat_G$ specifies a realisation for the observed events by assigning a functions for each variable node $X_i \in G$ such that $f_i(\mathrm{Par}_G(X_i), U_i) = (X_i = x_i)$ for some $x_i \in \mathrm{Val}(X_i)$ as determined by the influence of its parents and/or the influence of some unobserved effect. In addition $Feat_G$ determines the probability of each unobserved variable i.e. $\sigma(U_i)$ - each unobserved event is assumed to be independent of all others. We think of $Feat_G$ as providing the details of our causal intuitions regarding the relevant parameters for any causal relation amongst the nodes in G.

Some of the assumptions of this definition are questionable. For instance, why insist that we find a probability estimate for each unobserved event U_i? Is this probability estimate, subjective or objective? But of course, if we are to trust our causal intuitions then this is obvious. We ought to assign high probabilities if we think the unobserved features encode a crucial background condition. Suppose we wish to model the causal effect of striking a match, so we deliberately assume that the unobserved fact, that there is sufficient oxygen in the air. Our causal model might include this assumption, without presupposing it is directly causal, by simply assigning a high probability to the event without indicating any direct causal influence. The "unobserved parameters" are those which are assumed to hold with some or other degree of surety depending on the nature of the model. In short the unobserved variables usually encode those features which ensure that the causal relations turn out to be functional. The really important features of each model are the relations between the observed elements. With this definition in mind, we should be able to come up with a preference criterion over causal models.

Think of the world as containing all causal structures. When we seek to identify one or more of those causal relationships we propose a model of a causal structure which we claim is latent in the world. This proposal is to be preferred to all other proposals if we can see that all other structures can be interpreted as containing our proposed structural relation and a compatible probability distribution. In other words, it is the simplest model of this relation.

Definition (Latent Structure) A latent L = (G, O) where G is a causal structure and O is the set of observed variables $O \subseteq V$.

The idea is that in a latent structure we deliberately abstract somewhat from the entire set V of variables and only examine those we have observed (or are focused upon).

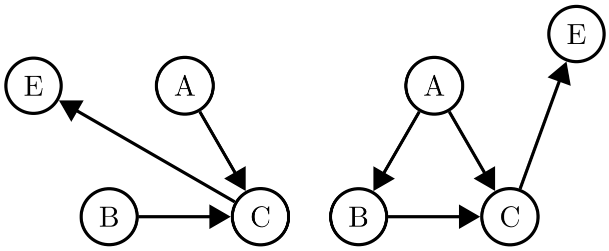
Definition (Structure Preference) One Latent structure L = (G, O) is preferred to another L'= (G', O') written, L $\preceq$ L' if and only if for every $Feat_G$, there exists a $Feat_{G'}$ such that the probability distribution over the observable events in our causal model based on L is equivalent to some probability distribution over observable events for the causal model based on L', written $\sigma_O(G, Feat_G) = \sigma_{O'}(G, Feat_{G'})$. In short, we wish to be able to recover L from any L' via a translation of conditional probability measure σ into a DAG.

Two latent structures are deemed equivalent if they are equally preferred. A related notion of *observational* equivalence can be defined for when two graphs share the same skeleton.[29] In any case, for a given set $\mathbb{L}$ of latent structures, we say that L is maximally preferred to L' in $\mathbb{L}$, written L $\preceq_{max}$ L' to all L' $\in \mathbb{L}$ iff L $\preceq$ L' for all L' $\in \mathbb{L}$.

[29] This is discussed in [73]pg19

Definition (Minimality) L is the minimal (read simplest) model available just when L $\preceq_{max}$ L' to all L' $\in \mathbb{L}$.

Ockham's razor states that we should always prefer the minimal model. Now we are in a position to use this criterion to identify genuine causal effect. Consider the following two models.



If the idea is to determine whether or not particular causal relations hold amongst the variable parameters of our theory, and we presume that simplicity is a reasonable criterion for theory choice, we should always take the model on the left to be superior to the model on the right. Is this too simple? What if A really is the cause of B? How is the model on the left better off for lack of true information? The issue at hand is what is the cause of D. The causal patterns resulting in B are to be considered in another set of model comparisons, wherein B is the child node. But again, the expansion of our causal models to account for the potential causes of B and D, can also be assessed for simplicity. This prompts the following definition.

Definition (Inferred Causation) Given a G compatible σ function, we say that a variable C *has a causal influence on variable* E if there exists a directed path from C to E in every minimal latent structure consistent with σ.

This definition allows us to recover a preference for the green-hypothesis over the grue-hypothesis on the basis that any time indexed colour change would require an extra causal mechanism to ensure that our perception of emeralds switches from green to blue. Whereas, accepting the green-hypothesis allows us to insist on one causal mechanism for all time thereby explaining the observed fact that emeralds are seen to be green with the most economy. Any σ-function compatible with the green-hypothesis on the basis of the accumulated evidence will by the nature of Goodman's argument be compatible with a grue-model. However, the latter is structurally exorbitant, and so rejected. This result shows that we can define a purely graphical criterion for the development of theory choice, and indeed Pearl argues for others. However we defer discussion of these for another time[30]

5.4 Causal Reasoning

The problem of identifying causal relations is two part if we accept that causal relations are identified as those for which we reach communal agreement and endorsement i.e if we think to conclusions of scientific method to be merited. On the one hand, each agent is obviously limited in their exposure to frequency based tests of causal hypotheses so their reports hold limited validity, and on the other hand people are prone in experimental settings to derive their own causal assumptions even if experimenters explicitly instruct the participants on which causal hypotheses should be assumed.[31] Of course, this final observation only supports our claim that causal reasoning takes

[30] For further discussion of the graphical criteria for model preference consult Pearl in [73]pg41-50 and pg77 - 85 or Glymour *et al* [40] Chapter 5. These criterion are further used to develop distinctions between the notions of genuine and potential cause and spurious association. In general the depth of Pearl's thinking on these issues is beyond the scope of this chapter. We hope simply to present a succinct summary of how his thought can be see to address the problem of evidentiary relevance.

[31] This interesting observation is reported in [1]pg624

place on the basis of an active causal theory. In so far as Pearl's Bayesian networks allow us to represent and test our causal theory - and the nature of the causal relation, he provides an ideal environment in which to make our causal assumptions explicit.

5.4.1 The Logic of Causal Dependence

We offer Pearl and Halpern's discussion of causal reasoning without frills. We shall point to some deficits after we have developed the theory. Ultimately these deficits are not terribly worrying.

Syntax and Semantics

The language of Pearl's counterfactual logic takes its cue from the Bayesian network setting. As such the the syntax is remarkably uncomplicated. All atomic formulas are of the following form: (X = x), and we allow boolean combinations of such claims. This is just to say that we can express the observations that some events have been realised. The more interesting claims we can express are of the following schematic form: $[\phi](\psi)$. This is meant to be understood as the claim that ϕ *is the actual cause of* ψ.

However, since causal dependence is a relation between events, the formula is usually slightly more involved. A typical claim will state that the event E = e is the causal result of the fact that some events $\mathrm{X}_1 = \mathrm{x}_1 ... \mathrm{X}_n = \mathrm{x}_n$ have been realised, i.e. (E = e) is caused if we fix $\mathrm{X}_1 = \mathrm{x}_1 \wedge ... \wedge \mathrm{X}_n = \mathrm{x}_n$, written $[\mathrm{X}_1 = \mathrm{x}_1, ... \mathrm{X}_n = \mathrm{x}_n](\mathrm{E} = \mathrm{e})$. So much for Syntax.

The real details come in at the level of the semantics. We call S = (U, O, Val) a *signature* of a causal structure, where U is the set of unobserved (exogenous) parameters, O is the set of all observed (endogenous) variables, and Val as before is the set of potential values with which each variable can be realised. We could assume as before that we are dealing with a Qualitative setting, so $\mathrm{Val(X)} = \mathrm{x} \in \{0, 1\}$, but for the moment this is not crucial. Similarly, a causal semantic model M = (S, $Feat_S$) is defined on top of our signature so that we can track the functional relationships between the events in O∪U. The only difference between a causal *semantic* model and a causal model (defined above) is that we do not yet specify a probability function over the variables in S and the functional relations are stipulated, not "read off" the graph. Evidently a graph can be recovered once the functional dependencies are stated. In other words we only require that our model be detailed enough to validate claims of the form: $[\mathrm{X}_1 = \mathrm{x}_1 ... \mathrm{X}_n = \mathrm{x}_n](\mathrm{Z} = \mathrm{z})$ so we need only ensure that there is a functional relation which ensures the fact that Z = z, just when we set $\mathrm{X}_1 = \mathrm{x}_1 ... \mathrm{X}_n = \mathrm{x}_n$. Thought of as a causal history of the world we might seek to abbreviate the notation. So $[\mathrm{X}_1 = \mathrm{x}_1 ... \mathrm{X}_n = \mathrm{x}_n]$ becomes $[\overrightarrow{X} = \overrightarrow{x}]$. Similarly, for the unobserved variables. So truth in a model is determined by whether an event is observed to be true modulo the functional specification of S, given the realisation of particular unobserved facts.

Definition (Truth and Satisfaction)

Atomics (M $\overrightarrow{u}$) $\models_S (X = x)$ iff $[X = x] \in \mathrm{O}_{Feat_S}$.

Negation (M $\overrightarrow{u}$) $\models_S \neg(X = x)$ iff $[X = x] \notin \mathrm{O}_{Feat_S}$.

Boolean (M $\overrightarrow{u}$) $\models_S (X = x \wedge Y = y)$ iff $[X = x] \in \mathrm{O}_{Feat_S}$ and $[Y = y] \in \mathrm{O}_{Feat_S}$.

Actual Causation (M $\overrightarrow{u}$) $\models_S [\overrightarrow{X} = \overrightarrow{x}](\phi)$ if the following conditions hold:

- **(AC1)** (M, $\overrightarrow{u}$) $\models_S (\overrightarrow{X} = \overrightarrow{x}) \wedge \phi$
- **(AC2)** There exists a partition $(\overrightarrow{Z}, \overrightarrow{W})$ of O with $\overrightarrow{X} \subseteq \overrightarrow{Z}$ and some setting $(\overrightarrow{x^*}, \overrightarrow{w^*})$ such that (M, $\overrightarrow{u}$) $\models_S \mathrm{Z}^* = \mathrm{z}^*$ for some subset $\mathrm{Z}^* \in \overrightarrow{Z}$ where:

(a) (M, $\overrightarrow{u}$) $\models_S$ [$\overrightarrow{X} = \overrightarrow{x^*}$, $\overrightarrow{W} = \overrightarrow{w^*}$]$\neg(\phi)$.

(b) (M, $\overrightarrow{u}$) $\models_S$ [$\overrightarrow{X} = \overrightarrow{x}$, $\overrightarrow{W^*} = \overrightarrow{w^*}$, $\overrightarrow{Z^*} = \overrightarrow{z^*}$]$(\phi)$ for all subsets $\overrightarrow{Z^*}$ of $\overrightarrow{Z}$

(AC3) $\overrightarrow{X}$ is minimal. That is to say, no subset of $\overrightarrow{X}$ satisfies AC1 and AC2.

The only interesting definition is the one for **Actual Causation**. Given the preceding discussion we should recall that the goal is to isolate the set of events in the model of our world upon which ϕ is causally dependent. Hence (AC1) needs to hold since otherwise we cannot even correlate the truth of ϕ with the proposed cause ($\overrightarrow{X} = \overrightarrow{x}$). Understanding (AC2) takes a little more effort. The idea is to recover the parental child relationship of the graph structures. So think of $\overrightarrow{Z}$ and $\overrightarrow{W}$ as defining the sets of ancestor nodes, and descendent nodes in a graph with respect to the ϕ-nodes. By stipulating that $\overrightarrow{X} \subseteq \overrightarrow{Z}$ we aim to isolate the parental nodes of the ϕ-nodes. Hence the conditions (a) and (b) are designed to ensure that the set $\overrightarrow{X}$ are the unique parental nodes. The (a) condition states that regardless of the value of the descendants, if $\overrightarrow{X}$ is realised in any manner distinct from the specification in (AC1), then ϕ will not occur. This condition is supposed to encode the counterfactual definition of causation indebted to Hume, but well defended by Lewis in his "Causation" paper where the idea is that X = x causes ϕ just when if X = x were not realised, then ϕ would not occur.[32] Allowing that the descendent-nodes $\overrightarrow{W}$ can be realised in any manner is tantamount to insisting that causation is not a transitive relation contra Lewis, who simply stipulates that causation *is* a transitive relation, distinct from the intransitive relation of causal dependence.[33] The irrelevance of $\overrightarrow{X}$ to it's grandchildren is further reinforced by the (b) condition which stipulates that $\overrightarrow{X}$ is alone sufficient to ensure the truth of ϕ regardless of how either the indirect ancestors, or eventual descendents of the ϕ-nodes come to be realised. Of course (AC3) is the obvious analogue of the simplicity constraint discussed with regard to theory of inferred causation.[34]

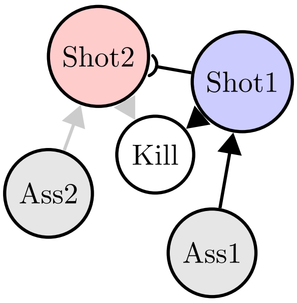
5.4.2 Causation or Causal Dependence

Why would we distinguish causation from causal dependence? Surely, the former is defined by the latter? If the latter is an intransitive relation, what could possibly prompt us to consider the former a transitive relation? I think the short answer is utility and economy of mental representation. Lewis argues for the transitivity of the causal relation by showing that even in the presence of competing causes, we privilege the simplest pattern of causation: Suppose that there is a chain of events whereupon an assassin is contracted to kill a target. Being diligent the contractor arranges for a fail safe. If the first assassin fails to shoot the target his backup will shoot the target. We want to say that the presence of one (not both) is the actual cause of killing, to do so we have to show how one causal sequence pre-empts the development of the other. Suppose the first assassin shoots, then the second assassin is pre-empted as his action is causally prompted by the inaction of the other. Depicted as follows:

[32] [1]pg632 - 638. His counterfactual theory of causation is very much disputed. Most significant criticisms comes from the fact that counterfactual conditionals are to be assessed on a possible worlds model where we have a similarity ranking over the worlds and A counterfactually implies B, iff the most similar A-worlds, are also B-worlds. Short of a tolerable definition of similarity any proposed ordering can be shown to question-begging with respect to the evaluation over some or other counterfactual claim by an underdetermination argument.

[33] We'll discuss this distinction below.

[34] The formal properties of a similar causal logic are discussed and elaborated in [44], where Halpern axiomatises the logic and proves soundness and completeness results in addition to developing some decision procedures.

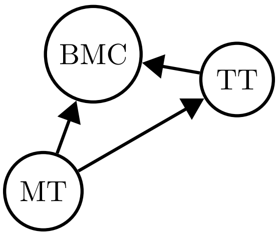


In the above picture we see that the first shot prevents the second shot, thereby rendering the killing causally independent of the second assassin, given the shooting of the first assassin. Since Lewis treats the causation relation as the ancestral of the causal dependence relation, he can infer that the first assassin caused the killing. In so doing he is able to distinguish an actual cause from a merely potential cause, but the event of killing does not causally depend on the assassin's existence in the same way that it does causally depend on the shot fired from the assassin's gun. This seems to suggest that we should distinguish transitive causation from intransitive causal dependence.[35]

In contrast Pearl and Halpern argue (following Hall) that causality is intransitive. They argue by counterexample. Let S = (O, U, Val) where O = $\{MT, TT, BMC\}$, U = $\{1stLD, 2ndAD\}$, and Val(2ndAD) =Val(1stLD) = Val(MT) = Val(TT) = $\{1, 0\}$ and Val(BMC) = $\{\spadesuit, \heartsuit, \infty, \bigstar\}$ Intuitively, we can think of this as a causal model for a treatment regimen for Billy the kid. He is either treated on monday, treated on tuesday, or treated on both days. We let M = (S, $Feat_S$) be such that S is as described, and $Feat_S$ is defined as follows:

$$
\begin{gathered}
f_1((\text{1stLD} = 1)) = (\text{MT} = 0)\\
f_2((\text{2ndAD} = 0)) = (\text{TT} = 0)\\
f_3((\text{2ndAD} = 1 \wedge \text{1stLD} = 1)) = (\text{TT} = 1)\\
f_4((\text{MT} =1 \wedge \text{TT} = 1)) = (\text{BMC} = \spadesuit)\\
f_5((\text{MT} = 1 \wedge \text{2ndAD} = 1)) = (\text{TT} = 0)
\end{gathered}
$$

We are attempting to find the causal chain which resulted in Billy's medical condition on Tuesday afternoon, where either Billy dies (♠) after recovery due to double dosing, remains sick (∞) in perpetuity due to no treatment, recovers (♡) after sickness due to late treatment, or is happy and healthy (★) due to early treatment. By assumption two doses of medicine are lethal. However, if the first doctor is lazy then there is no treatment on Monday, and if the second doctor is attentive, there will be treatment on Tuesday, if there has been no treatment on Monday. So, if causation is transitive, then we should be able show that MT causes TT, and since TT causes BMC, we know that MT causes BMC too. We depict the situation as follows:



Now we can develop a counterexample by using the definition of **Actual Causation**. Given our set up, it's clear to see that MT = 1 is a cause of BMC = ★ since (AC1) is met trivially. Take $\overrightarrow{W} = \emptyset$, then changing the value of MT = 0, ensures that we can never have an early recovery, thereby satisfying (AC2a) and similarly for (AC2b). Of course (AC3) holds because there is no

[35] Pearl and Halpern discuss an analogous case that is formalised in their logic of causal dependence in [47]. We defer the elaboration in favour of focusing on another application of their system.

other sufficient cause for BMC $= \bigstar$ contained in the model at all. By similar reasoning we can check that TT $= 0$ causally depends on MT $=1$, and TT $= 0$ causes (BMC $= \infty \vee$ BMC $= \heartsuit \vee$ BMC $= \bigstar$) i.e. Tuesday's treatment is the cause of Bill's continued existence whatever state that induces. But MT $= 1$, is not the cause of Bill's continued existence, since MT $= 1$ does not cause (BMC $= \infty \vee$ BMC $= \heartsuit \vee$ BMC $= \bigstar$) because (AC2a) fails. If we change MT $= 0$, it still follows that (BMC $= \infty \vee$ BMC $= \heartsuit \vee$ BMC $= \bigstar$) by the realisation of TT.[36] So the relation of **Actual Causation** is not transitive.

Criteria of Adequacy: Metaphysical or Definitional

There is a familiar debate over the reality of necessary connection. Hume argued no such connections exist and that the misguided insistence on causal law was an error prompted by observations of constant conjunction. In a similar vein Quine can be seen as arguing that modal facts hard-code information about definitional connections rather than essential connections. Quine argues as follows: Suppose "9 is essentially composite" i.e. that there is a true modal fact which stems from the *de re* modal properties of an individual. Observe similarly that "9 = the number of planets", but then by Leibniz's law we get the following substitution instance "The number of planets is essentially composite", but this is false. By the truth of the second premise and Leibniz's law we should infer the falsity of our first premise and thereby reject the cogency of *de re* modal predication. Or so goes the argument.[37]

In so far as causal truths encode modal information, we might wonder whether causal truths are merely definitional, or in fact metaphysical. Let us try to determine whether causation is a transitive or non-transitive relation. This is surely a substantive question deserving of an answer one way or the other. First we reflect on the considerations which aim to ridicule the metaphysical reading.

Construed charitably Quine's argument is unsound, but hides an interesting issue. Treated under a *de re* reading, the conclusion that the number of planets is necessarily composite, sounds outlandish but it is not obviously cogent, so certainly not obviously false! Thus the conclusion cannot be used as a *reductio* of the assumption.[38] We are better off to take Quine's argument as an indicator of the extreme epistemological difficulty faced by the aspiring metaphysician. Forced to reckon with *prima facie* incomprehensible claims the metaphysician braves an unwelcoming environment every time they open their mouth to speak - due primarily to the unknowable nature of their pronouncements. It is then natural to think that Quine was advocating the view that only a *de dicto* reading of such modal claims was plausible. In such a case, the conclusion doesn't follow from the premises, but premises can be added to either corroborate or deny such a definitional connection. Hence modal claims are forever unknowable and we may choose as we like *a la* the conventionalist.

However, we have already argued that there is little reason to adopt structural conventionalism. The unknowability of a particular claim follows only if knowledge is defined as truth over all logically possible worlds, rather than a limited subset of those worlds. Quine's argument does not significantly change this conclusion. However, it does bring into stark relief the poor reasons for adopting conventionalism. In particular, if the problem of treating causal claims as *de re* facts, stems from the epistemological difficulty then Pearl's suggestions for identifying causal relations put paid to Quinean style objections. Our procedures of identification are defeasible but this alone is not a reason to suggest we treat causal claims as implicitly definitional. Hence, lacking a

[36] In addition this result shows that causality is not closed under *right weakening* because we know that MT $= 1$, ensures that BMC $= \bigstar$, so if *right weakening* held, we would know that BMC $= \bigstar \vee$ BMC $= \heartsuit$ which does not hold by the definition of actual cause.

[37] See Quine's "Three Grades of Modal Involvement" in [74]

[38] This point is also made by Frank Jackson in [43]pg259

compelling argument to the contrary we will proceed to assess the adequacy conditions for causal claims as being mandated by empirical corroboration - and by extension, our results should be thought apt for ontological generalisation.

With this in mind, we can ask whether Lewis' contention that causality is a transitive relation has a better motivation than Pearl and Halpern's argument for the contrary claim? First note that Lewis' preemption argument can at best be seen as supporting the transitivity of causation by appeal to (a) a preference for simpler causal explanations and (b) felicity with causal intuition. Pearl and Halpern's definition of **Actual Causation** is intended to do exactly the same. A preference for simplicity is encoded by the insistence upon (AC3), and our causal intuitions are incorporated with the dynamic nature of the functional relationships our models instantiate. But who is wrong and why? These are poor questions.

5.4.3 Explanation and Causation

We have a case of conflation. Assuming that the object of the causal theorist is to identify the metaphysical relation of causation and its formal properties by examining instances of causation assumed in our best scientific practice, then we should not be particularly worried about recovering the"folk" notion of causality because "the folk" tend towards transient characterisations of such relations, as suits their immediate preferences.[39] If we observe that "the folk" are fickle people, any model of the causal relation which sought motivation in the whims of the folk would be motivated only until the winds change. Yet the test of a causal theory is its intuitive fit with our causal reasoning. There seems to be a tension here. Either we have conflated folk intuition with the properly speaking metaphysical relation of causation, or we have mistaken the latter for the former.

On the one hand Lewis seems to be involved in an exercise of conceptual elaboration wherein he attempts to explicate the "folk" notion of causation. Simultaneously prescriptive and arguably descriptive. His insistence on the transitive nature of the causal relation is plausibly motivated by the inattentive but prevalent usage of reasoning by causal chains. On the other hand Pearl and Halpern argue that the actual relation of causation cannot be transitive by appealing to similar intuition. So how might we develop a theory of explanation atop the theory of causation? Should we? Consider Pearl and Halpern's attempt to define the explanatory relation on top of the **Actual Causation** relation:

Definition (Causal Explanation) Given a causal model M, we say that the pair (ψ, $\overrightarrow{X} = \overrightarrow{x}$) is an explanation of ϕ relative to our epistemic context $\mathcal{K} = \{\overrightarrow{u} \mid \overrightarrow{u} \sim \overrightarrow{u_*}\}$ if the following conditions hold

(EX1) (M, $\overrightarrow{u}$) $\models \phi$ for each context in $\mathcal{K}$. In other words, we are trying to explain an acknowledged fact ϕ.

(EX2) [$\psi \wedge \overrightarrow{X} = \overrightarrow{x}$] is a *weak cause* of ϕ in (M, $\overrightarrow{u}$) just when (AC1) and (AC2) hold, but (AC3) does not necessarily hold, while M $\overrightarrow{u} \models (\overrightarrow{X} = \overrightarrow{x})$ and M $\models \psi$.

(EX3) $\overrightarrow{X}$ is minimal i.e. no subset of $\overrightarrow{X}$ satisfies (EX2) in a situation where ψ is valid.

(EX4) (M, $\overrightarrow{u}$) $\models \neg(\overrightarrow{X} = \overrightarrow{x})$ for some $\overrightarrow{u} \in \mathcal{K}$ and (M, $\overrightarrow{u_*} \models (\overrightarrow{X} = \overrightarrow{x})$ for some $\overrightarrow{u_*} \in \mathcal{K}$

The idea behind this definition is closely related to the concept of knowledge defined over possible worlds, but we add the proviso that those possible worlds are such that we know particular causal information of the form $[\chi](\tau)$ to be encoded by the formula ψ. Thereby ensuring that each possible context $\overrightarrow{u}$ is such that we know ψ, so that we may determine whether or not ($\overrightarrow{X} = \overrightarrow{x}$)

[39] Think here of reports of causation in politics.

is an explanation of ϕ relative to our structural knowledge and the observed state of the world. We then test to see whether $(\overrightarrow{X} = \overrightarrow{x})$ is a weak cause of ϕ given ψ, if so we further check that it is the simplest weak cause of ϕ. This is quite a reasonable process by which to discover explanations.

The condition (EX4) is to be maintained to avoid making our explanations trivial. We need to presume that $(\overrightarrow{X} = \overrightarrow{x})$ is not already known and that it is also possibly true. These are stated to be crucial constraints on an explanation. Arguably, this condition is contrary to intuition. For surely, I may find out that some presumed fact counts as an explanation for a novel observation? Hence you might expect that $(\overrightarrow{X} = \overrightarrow{x})$ could indeed be known, and nevertheless serve as an explanation for ϕ. Another problem arises if we look to natural language explanations. Consider the claim that "Victoria took a vacation in the Canary Islands" as a candidate explanation for her current tan. Intuitively, this is sufficient, however on the above model it would fail. For Pearl and Halpern's definition to work we would need to add the conjunction "Victoria took a vacation in the Canary Islands *and* it was sunny". Both these problems suggest the nature of divergence between intuition and the model of explanation above: the model of explanation is model of a process not of a relation. Pearl and Halpern insist on (EX4) because the process of discovery is supposed to motivate the proposed explanation! If we assume $(\overrightarrow{X} = \overrightarrow{x})$ is known than the correlation with ϕ is easily achieved and this can be done for any arbitrary connection you care to make. Suppose the sun rises and the cock crows everyday, note that neither causes the other. In this way (EX4) can be seen to block quus-like cases of deviant connection. In everyday speech we treat explanatory claims as encoding a simple relation between two clams. Explanatory claims are usually short hand for a more detailed story which we seek to imply by stating "χ explains τ". Taking the above model of explanation seriously, involves adopting the view that the process of explanation is the process of discovering the underlying mechanisms which relate χ to τ.

Grant our premise that explanation is the process of discovering connections, and reporting said connections. Now we can respond to Lewis' preemption argument. While Lewis argues for the transitivity of the causation relation because he identifies causal influence with the ancestral of counterfactual dependence, we should note that counterfactual dependence is alone an insufficient criterion for identifying causal influence, or rather, it is only part of the conditions required to identify causal influence. Think here of the Suppression test as giving rise to examples of interrupted causal chains, thereby undermining preliminary estimates of causal influence based solely on intuitions regarding counterfactual dependence. When coupled with other criteria we can see (as above) that there are cases of non-transitive causation. This concludes the debate to my mind. Pearl and Halpern's counterexample is compelling, and I think decisive. Causality is not a transitive relation. We might understand Lewis' dogmatic insistence to the contrary as indicative of an understandable confusion between causality and a simplistic notion of causal explanation, wherein the latter is conveniently transitive for ease of recall. But such a theory of explanation pays insufficient attention to the facts.

So far so good. However the model is not entirely problem free. The main issue relates to the nature of the causal explanations which emerge when based upon an intransitive relation of causal dependence.

Example (The Loan shark)

Intuitively if you cut off your own finger (FC), you are not the cause of it being fully functional (FF) a month later. However, we consider the idea that had you not severed your own finger, then Larry the loanshark would have removed your finger (LC) and thrown it in the rubbish because he was waiting (LW) for too long on your repayment of an old debt. As such, your finger would never be sown back on, and hence never fully functional i.e. if $\text{FC} = 0 \wedge \text{LC} = 1$, then $\text{FF} = 0$. When you sever your own finger you are immediately rushed to the hospital bypassing Larry. In such a model we can show counter intuitively that the act of cutting off your own finger is the **actual cause** of it being fully functional because FC = 1, then FF = 1, since LC = 0. The example can be fleshed out more formally if required, but the problem should be clear.

How do we decide whether or not to include the Larry contingency in our causal model? Is this a species of underdetermination? Certainly. Is this detrimental to the argument we have been developing? Not particularly. While we seek to undermine certain kinds of underdetermination argument by including structural information in our model we do not pretend to dissolve or resolve all underdetermination issues. For their own response Pearl and Halpern seek to ignore such problems by imposing a plausibility ranking over events so as to exclude extraneous contingencies. Presumably the Larry-type of contingency can be reasonably excluded from the model if we know something about your gambling habits. In effect, this becomes a modelling issue, and the faithfulness of the model is to be assessed with respect to your broader knowledge or the situational parameters. So we just need to unravel the causal sequence to determine whether LW = 1, given the appropriate considerations about your gambling. This is possible at the meta-level of model construction but the way in which counterfactual conditionals are elaborated in Pearl and Halpern's setting precludes the expression of backtracking-causal counterfactuals in the object language. This point is made forcefully by Rips[40]. Whether this is a real deficit of the theory depends on what you expect from your model of causation, but felicity with all species of counterfactual claims is not a necessary condition of causal reasoning.

5.5 Conclusions

In this chapter we have sought to demonstrate the procedure discussed in the previous chapter. Our desire was to examine the relation of causal independence so as to develop an express theory of causal explanation. Throughout we sought to motivate this investigation by displaying the role causal information plays in our theory preference. In particular we reiterated the idea that the development of a fine-grained causal theory can dissolve the grue-paradox. So we conclude this chapter having done as desired. However, there remain some problems and infelicities between our model of explanation and our inherent intuitions. Ultimately, these problems relate to the manner in which we are to fix the unobserved parameters U_i in our models of causality.

This is to be expected. As Goodman initially argued we cannot import assumptions about causal relevance without somewhat prejudging some or other open question. However, given that the specification of these parameters is required for investigation into the causal structure of the world, we can argue that such specifications are only appropriately motivated by pragmatic considerations regarding the nature of the question under investigation. These assumptions will be defensible in each instance, and cannot be motivated *a priori*. Modelling is task orientated and as such, the features of each model will respect the traditions of the task where those traditional assumptions are merited. Halpern and Hitchcock take this to indicate that causal judgements are in some sense subjective.[41] However, we also find efforts to support these pragmatic choices by incorporating default assumptions into their causal models.[42]. To mandate any default rule we must resort to the citation of some public criterion of assessment, thereby making said rule at least not entirely subjective. Evidently there are further issues to explore but minimally we take it that the role of structural information has to play in explanatory reasoning, and by extension epistemological models, should now be clear.

On this note we conclude the chapter. In the next (and final) chapter we briefly sketch and contrast a theory of explanation-as-information-update against some of the traditional theories, and

[40] [1]pg622-623

[41] In [46]

[42] See for instance [48]

then survey and respond to some objections to the general project to base a theory of explanation on theoretical dependency structures.

Chapter 6

Arguably, arguably true: Objections and Replies

> *Sur quelque préférence une estime se fonde, Et c'est n'estimer rien qu'estimer tout le monde.* - Moliere in *Le Misanthrope*

6.1 Introduction

In this chapter we aim to cover three tasks: (i) we shall abstract from our foregoing discussion and draw some conclusions regarding the process of explanation (ii) we shall consider broad objections to the species of scientific and metaphysical realism we have been advocating throughout this thesis and (iii) we summarise some competing theories of explanation and argue that ultimately our approach is both superior and complimentary.

We shall first develop the notion of explanation as an update operation in a dynamic epistemic logic. These considerations shall be at best suggestive, because full formal elaboration of the details is beyond the scope of this thesis.[1] With this background in mind we shall vindicate our approach to the theory of explanation by a comparative contrast with an existing theory of explanation. We then defend our theory against some reasonable objections. The aim is to show that our approach to dealing with underdetermination problems is well supported because attempts to undermine our approach lead to incoherence at worst, or at best suffer from plausibility problems. We draw primarily from the work of van Fraassen.[2] In particular we shall defend the theory we have elaborated from van Fraassen's objections relating to the criteria of adequacy on a scientific theory.

6.2 The Two Axes: Systematic concerns

Abduction problems are best seen as a search for an explanation. This prompts certain adequacy conditions on an solution to such problems. There are two axes on which an explanation is to be assessed. On the one hand, we should determine that the explanation is true, and on the other we need to determine that the explanation is good. What precisely makes an explanation good appears to depend wholly on the context, i.e. who asked the relevant why-question, what they know (or believe) about the topic, and whether or not your answer was comprehensible given these limitations. On the one hand we can see this as a distinction between a pragmatic and idealised approach to explanation. The pragmatic approach will emphasise the importance of effective

[1] In particular we do not make explicit the connections between the dependency logics, and justification logics with awareness functions *a la* the work of Velazquez in [84], although much of this connection underlies the suggestions we shall make. Such explicit connections are beyond the scope of this thesis.

[2] Especially [32] and [33].

communication with the intent to induce comprehension and understanding. Whereas the ideal approach will seek to pursue correct (or true) explanations. Under this lens it should be clear that our theory of explanation has only been developed to cater for the demand that an explanation is true. We have not developed a theory adequate to meet the demand that an explanation be good. The reason for this deficit is that we have not extended out discussion to a multi-agent setting and it is difficult to establish the properties which make an explanation good without factoring for the effect of their deployment in a multi-agent setting.

We shall now borrow some ideas from contemporary work in Dynamic multi-agent epistemic logic to indicate how we can extend our theory to render our true explanations, good.

6.2.1 Explanatory relevance and Justificatory update

Note that we have observed a number of species of explanation of differing strengths which are apt to induce different responses in any agent who accepts such an explanation. So we should be concerned to indicate this subtlety in the patterns of acceptance performed by rational agents. First things first. We need to be able to say when an agent *accepts* an explanation. Formally we can make this a test on the agents information states, if an update with a proposition effects no change in an agents information state, then we can see that an agent accepts (i.e. knows, believes or has a justification for) the proposition. The acceptance of the appropriate information is a precondition for further information to function as an explanation for the previously accepted claim. We can make this a bit more formal if we borrow from Veltman's update semantics[3]

Definition (An Update System) U = $< \mathcal{L_U}, \sum [\,], >$ where $\mathcal{L_U}$ is the smallest set of sentences formed by recursive construction i.e.

$$\psi ::= \text{p} \mid \neg\psi \mid \tau \wedge \psi \mid \Diamond\psi$$

and $\sum$ is a set of information states, one for each agent i.e. a set of possible worlds consistent with all the agent accepts, wherein each world can be thought of as a function assigning values to the propositional variables. Finally, [ ] is an update function performed on an information state with a particular proposition. For example let Υ_i in $\sum$ be information state, then we write Υ_i[p] where some agent i wishes to accept p into their information state Υ.

Definition (Acceptance) A sentence ϕ is accepted in an information state Υ iff $\Upsilon[\phi] = \Upsilon$ i.e $\Upsilon \models \phi$.

Evidently this notion of acceptance is strongly tied to the notion of knowledge, but we might consider defining weaker standards of acceptance such that we update in such a way that the new information comes to be believed rather than known. For the moment, this will suffice.[4] The update function as elaborated here operates on the information state but is not part of the syntax of $\mathcal{L_U}$, so the effect of an update is not, so to speak, known by any of the agents we model. Both Jelle Gerbrandy and Fernando Raymundo Velazquez Quesada offer suggestions about how to include awareness of the effect of an update within the information states of our agents.[5] In any case, we can still calculate the effect of an update on any information state as follows:

[3] [94]

[4] It's important to note that the update operations we are about to define supposes nothing about the structure of the relevant information states. So we may easily consider characterisations of an information state as determined by justification logic also.

[5] In [38] and [84]

Definition (Update Effects)

$$\Upsilon[\mathrm{p}] = \{w \in \Upsilon \mid w(p) = 1\}$$

$$\Upsilon[\neg\phi] = \Upsilon \setminus \Upsilon[\phi]$$

$$\Upsilon[\phi \wedge \psi] = \Upsilon[\phi] \cap \Upsilon[\psi]$$

$$\Upsilon[\Diamond\phi] = \begin{cases} \Upsilon & \text{if } \Upsilon[\phi] \neq \emptyset \\ \emptyset & \text{if } \Upsilon[\phi] = \emptyset \end{cases}$$

In addition if we wish to update with a sequences of sentences, we insist that $\Upsilon[\phi_1; ...\phi_n] = \Upsilon[\phi_1]...[\phi_n]$. Most clauses are straightforward, the modal claim operates like a consistency check for ϕ with our available information.

So far so good. But how do we incorporate updates with defeasible default rules or evidentiary assumptions? We need to extend the language $\mathcal{L}_{\mathcal{U}}$ and distinguish between the factual (propositional) information of an agent and the rules of expectation which they adopt. We might, for instance assume, as before, that we have an normality ordering on the worlds in our information state such that if $\Upsilon \models \phi \rightsquigarrow \psi$, then we might expect $\Upsilon[\phi] = \{w \in \Upsilon \mid w(\phi) = 1 \wedge \mathrm{MAX}_{\preceq^{\phi}_{N}}(\psi) = 1\ \}$. Alternatively we can adapt the evidence function $\mathcal{E}$ from the justification logic setting. Treat testimony as a form of evidence with which we update our information state, then we might accept the claim that ϕ serves as explanatory evidence for the occurrence of ψ. So in terms of the question why-ψ, we could profitably answer the question with ϕ just when we know that our interlocutor accepts either the default rule $\phi \rightsquigarrow \psi$, or acknowledges that ϕ is evidence for ψ i.e (e_{ϕ}:ψ).[6] In effect this stipulation would have us treat default rules (or evidentiary connections) as encoded in the presuppositions of our interlocutor. If no such presupposition exists then our testimony gets treated as true, but we do not ensure that our conversational partner appreciates the relevance of ϕ to ψ or that ϕ comes to count as an explanation for anything. Another helpful distinction is that of implicit and explicit belief and knowledge. We might want to distinguish *a la* Velazquez[7] between explicit and implicit beliefs and knowledge, so that we can think of an explanation as an information update in which some implicit belief $\mathrm{B}^{Im}\phi$ is rendered explicit $\mathrm{B}^{Ex}\phi$. These ideas are straightforward but we defer full elaboration for later work.

The real problem often relates to convincing others that ϕ is indeed relevant to ψ. Our announcement of such a claim should not be sufficient to induce this belief in our conversational partners. Rather the source by which each agent derives their notion of evidentiary relevance must be independent of their sources for simple testimonial evidence. The purpose of this thesis has been to suggest that by developing beliefs and knowledge in the structural relations of dependence between events, each agent supplies their own relevance relation which constrains the formation of explanatory answers. You cannot simply announce that e_{ϕ} is evidence for ψ i.e. $[\mathrm{e}_{\phi} : \psi]$ and expect the behaviour of others to conform to this rule. In a sense, all arguments about evidentiary relevance have to be resolved by appeal to the phenomena in question.

6.2.2 Dynamic Update and Explanation as Public Announcement

But at any point where we have resolved our views about evidentiary relevance we are still left with the open question regarding how this information should effect our beliefs and actions. As such we need to distinguish between the types of model changing operation. We canvas three options from

[6] This is subtly different from the usage of $\mathcal{E}$ in the justification logic setting. There the evidence function was explicitly tied to belief, whereas here we are treating it as determining evidentiary relevance from which we can induce belief or knowledge depending on the nature of the evidentiary relation it tracks.

[7] See [84]

the literature and suggest that we should think of explanation as an analogous form of update with more complicated restraints on the preconditions.[8]

First formulated with respect to the logic of public announcement[9], the following operations were thought of as public announcement of information, potentially as answers to a particular question. Each answer is more or less informative depending on the reliability of the source of the announcement.

Learning with Certainty ($!\phi$) The idea here is that you are informed with absolute certainty that ϕ is true. In effect ϕ is the answer to the question Q $= \{\phi, \neg\phi\}$. Since $!\phi$ is an update operation on our model we define the effect of learning ϕ as the operation which deletes all the non-ϕ worlds from our information state Υ. Formally $\Upsilon[!\phi] = \Upsilon^* = \{w \in \Upsilon \mid w \models \phi\}$ where the Kripke (or plausibility) relation over Υ now holds only for the ϕ-worlds remaining in Υ^*. This type of learning may be treated as announcements from an infallible source such as an oracle or more direct observations of Nature.

Developing Preference ($\Uparrow \phi$) Assuming that oracles are few and far between we should observe there are other methods of learning. Indeed, we might learn to develop a preference. Assume that there is a plausibility relation on Υ. When faced with the question Q $= \{\phi, \neg\phi\}$, if we develop a preference for ϕ then the $\Uparrow \phi$ operation moves all the ϕ worlds up the plausibility ranking such that no ϕ-world is less plausible than a $\neg\phi$-world. Formally $\Upsilon[\Uparrow \phi] = \Upsilon$ such that w is more plausible than w', where both model ϕ or both model $\neg\phi$, or w is more plausible than w' and w models ϕ and w' models $\neg\phi$. This type of update might be best thought of information from a trusted, but not infallible source.

Barely Believed Information ($\uparrow \phi$) The idea is that we have reason to believe ϕ, but it is not sufficient to make all ϕ worlds more plausible than the $\neg\phi$-worlds. We promote only the $\text{MAX}^{\phi}_{\preceq}$-worlds to the $\text{MAX}_{\preceq}$-worlds, so that ϕ becomes tentatively believed. However, unlike the $\Uparrow$-operation, if the ordering changes again we are unlikely to retain our belief ϕ. Formally, $\Upsilon[\uparrow \phi]$ $= \Upsilon$ such that w is more plausible than w' iff either w $\in \text{MAX}^{\phi}_{\preceq}$ or w $\preceq$ w' in Υ.

So in analogy with these varying types of upgrade operation we wish to insist that an explanation is a type of public announcement which induces a change of an agent's doxastic-epistemic state. However, we also should note that for an explanation operation to occur the agent must satisfy certain preconditions so that the dynamic effects of the announcement will appear reasonable.

6.2.3 Explanations: Specifying the Preconditions and Postconditions

On the one hand we need to specify that each question is an abductive problem with more or less elements in the contrast class. A solution to an abductive problem must be one of the elements of the contrast class considered possible by the agent. Hence, given that the contrast class partitions the information state Υ, a solution is one of the cells of that partition.

We may distinguish between single answer questions and multiple answer questions as those in which either the partition over Υ contains mutually exclusive answers, or compatible answers. This may be thought of as a difference in exactitude which governs different domains of questions and answers. Single answer questions are asked in a domain where questions can (and should) be given an exact answer. Multiple answer questions encode a kind of laxity about what constitutes

[8] In the following discussion we draw on the work of van Benthem, and Baltag and Smets in [92] and [10] respectively.

[9] cf [93]

a relevant answer to a particular question.

Belief in the relevance of the *explanans* is a minimal condition for an explanation-update to be triggered. But we can distinguish between implicit and explicit beliefs. In particular we need to determine whether our belief in the relevance of ϕ for evaluating ψ is implicit, or explicit - and what precisely this should mean when we update our information state with $\Upsilon[\phi]$. Let $\mathrm{e}_{\phi}{:}\psi$ denote the claim that there is evidence e for ψ and e $:= \phi$. Should we insist that receipt of an explanation renders an implicit belief in $\mathrm{e}_{\phi}{:}\psi$ to an explicit belief? Better, should we distinguish between different species of explanations and their impact on our beliefs? While causal evidence might be apt to induce knowledge, evidence of metaphysical dependence might be thought to induce belief at best (assuming an inclination for naturalism). The idea here is that we ought to distinguish between the sources of our explanations, to determine how to expand our beliefs and knowledge. A similar motivation runs through Pagnucco's discussion of abductive expansion functions:

> In abductive expansion... the concern is to determine which beliefs should be incorporated into the current epistemic state using an abductive strategy to identify the appropriate expansion given new information.[10]

We suggest that the correct input for an abductive solution is to be taken from our best theory of the appropriate dependency relation. In short, we should specify that $\mathrm{e}_{\phi}{:}\psi$ holds just when we have oberved some sense in which ψ depends on ϕ. However, we have not specified the nature in which an abductive solution enacts an expansion of our doxastic or epistemic state. Consider this a first effort

Predictive Explanation for Q

Let ξ_{ψ}^{pred} :ϕ be the update operation on Υ with a predictive explanation. We say that ϕ is a predictive explanation of ψ just when the preconditions (a) the ϕ-worlds are a cell in the Q_{ψ} partition of the state space, and (b) when we implicitly believe ϕ is relevant to ψ, induce the postcondition (c) that the inclusion of ξ_{ψ}^{pred} :ϕ in our information renders explicit our belief in this relevance and ensures that this connection holds in fact i.e. that ψ really does depend on ϕ, if this operation retains the consistency of our information state.[11] The idea is that an explanation will never be accepted if it contradicts our current information. This species of predictive explanation might be thought to be appropriate for causal explanations. We might also define other explanatory updates analogously.

Full Explanation for Q

Let $\xi_{full}^{\psi}{:}\phi$ be an update operation on Υ with a full explanation. We say that ϕ is a full explanation of ψ just when the preconditions, (a) the ϕ worlds are in the Q_{ψ} partition of the state space, and (b) ϕ is implicitly believed to be relevant to ψ, induce the postcondition (c) that we come to know explicitly the connection between the two propositions, and furthermore it follows that all other potential answers in the partition Q_{ψ} become false if the operation preserves the consistency of our information.[12] The idea here is that a fully explanatory answer to a why-ψ claim, is apt to motivate the abductive inference from ψ to ϕ because the latter serves as evidence for the latter, and excludes all other candidate answers, just when such an update does not render our information state inconsistent. So clearly this is a rare kind of update, more likely we receive weaker explanations.

[10] Quoted in [71]pg137

[11] Formally we might say that $\Upsilon[\xi_{pred}^{\psi}{:}\phi]$ iff (a) ϕ-worlds $\in \mathrm{P}_{Q_{\psi}}(\mathrm{W})$ and (b) $\Upsilon \models \mathrm{B}^{Im}(\mathrm{e}_{\phi}{:}\psi)$ and (c) $\Upsilon[\xi_{pred}^{\psi}{:}\phi] = \Upsilon[(\mathrm{B}^{Ex}\phi \leftrightarrow \mathrm{B}^{Ex}\psi) \wedge (\mathrm{B}^{Ex}\neg\phi \leftrightarrow \mathrm{B}^{Ex}\neg\psi), (\mathrm{e}_{\neg\phi}{:}\neg\psi), (\mathrm{e}_{\phi}{:}\psi)]$ if $\Upsilon[\xi_{pred}^{\psi}{:}\phi] \neq \emptyset$.

[12] Formally, we might offer the following definition: $\Upsilon[\xi_{full}^{\psi}{:}\phi]$ iff (a) ϕ-worlds $\in \mathrm{P}_{Q_{\psi}}(\mathrm{W})$ and (b) $\Upsilon \models \mathrm{B}^{Im}(e_{\phi} : \psi)$ and (c) $\Upsilon[\xi_{full}^{\psi}{:}\phi] = \Upsilon[(\mathrm{K}^{Ex}\psi \rightarrow \mathrm{K}^{Ex}\phi), (\mathrm{e}_{\neg\phi}{:}\neg\psi), (\mathrm{e}_{\phi}{:}\psi), \neg\bigvee(\tau)\forall\tau \in \mathrm{P}_{Q_{\psi}}$ where $\tau \neq \phi]$ if $\Upsilon[\xi_{full}^{\psi}{:}\phi] \neq \emptyset$.

Weak Explanation for Q

Let $\xi^{\psi}_{weak}{:}\phi$ be an update operation on Υ with a weak explanation. We say that ϕ is a weak explanation of ψ just when the preconditions, (a) the ϕ worlds are in the Q_{ψ} partition of the state space, and (b) ϕ is implicitly believed to be relevant to ψ, induce the postcondition (c) that we come to explicitly believe that ψ normally entails ϕ and the latter is evidence for the former if this operation retains the consistency of our information state.[13] This species of explanation includes update with a default rule, rather than a strict material implication. The idea is that such normality assumptions are appropriate in more speculative perhaps metaphysical explanations. We tentatively suggest this is the kind of explanation properly attached to metaphysical explanations based on the grounding relations. There are as many different expansion functions as there are dependency relations, and each such expansion characterises a different stripe of explanation. So the different uses of predictive and non-predictive explanation will reflect a difference between the source of the explanations and the nature of the dependence claims.

6.2.4 Common Knowledge: Controversial Explanations

Justification games[14] give us the clue that there are constraints on what counts as a justification, and where this counts. Similar observations can, but rarely are, made of explanations. Hence the contextual constraints on an explanation are less obvious, or at least less entrenched. For any given group assume some set of constraints holds.

Let G be a group of agents, then for each agent has an information state Υ_i, and so for every agent $i{\in}\mathrm{G}$, we have a predicate K_i for each agent in G. So we define an operator $\mathrm{E}_G\phi := \bigwedge_{i\in G}\mathrm{K}_i\phi$, which states that everybody knows ϕ. Then common knowledge in terms of $\mathrm{E}_G\phi$.

$$\mathrm{C}_G\phi := \textstyle\bigwedge_{n=0}^{\infty}\mathrm{E}^n_G\phi$$

The semantics for these two operations are of significant interest, since E_G is defined in terms of the set of all agents i.e. the union of the equivalence relations indexed to each agent in G, whereas C_G is defined on the transitive reflexive closure of the E_G relation.[15] With these definitions in place we can fix what is taken to be common knowledge in the group. Better still we can fix the type of explanations which are commonly known to be applicable in the group. It is an entirely non-trivial task to specify how common knowledge interacts with awareness sets and justification logic, but intuitively we might think to specify a group setting in such a way that an explanatory update is uncontroversial if the explicit information induced by such an update is already accepted by each member of the group and it is common knowledge that such information is uniformly accepted. An explanation is controversial otherwise.

Pursuing the analogy with a justification game we suggest that every controversial explanation in a group poses a resolution problem where each faction in the group must mount a defence of the of their explicitly conflicting assumptions, thereby prompting a justification game. Whether we conceive of an explanation as a move in a justification game or justification as a move in an explanation game, the same parameters must be addressed i.e. we must have mechanisms to defend (a) the construction of a question-partition, and (b) the generation of implicit beliefs, and (c) a mandate for choosing when (i) weak explanations, (ii) predictive explanations or (iii) full explanations are appropriate. We suggest that the appropriate defence of such assumptions rests with the nature of the dependence between the *explanandum* and the *explanans* and as such our focus on elaborating and carefully specifying the nature of these dependency structures is well motivated by the role they implicitly play in our justificatory or explanatory practices.

[13] Formally $\Upsilon[\xi^{\psi}_{weak}{:}\phi]$ iff (a) ϕ-worlds $\in \mathrm{P}_{Q_\psi}(\mathrm{W})$ and (b) $\Upsilon \models \mathrm{B}^{Im}(e_\phi : \psi)$ and (c) $\Upsilon[\xi^{\psi}_{weak}{:}\phi] = \Upsilon[(\mathrm{B}^{Ex}\psi \rightsquigarrow \mathrm{B}^{Ex}\phi)$, $(\mathrm{e}_\phi{:}\psi)]$ if $\Upsilon[\xi^{\psi}_{weak}{:}\phi] \neq \emptyset$

[14] As discussed for instance by [61]

[15] For more details see [93]pg30-40

6.2.5 Belief Revision for Identifying Explanations

To emphasise this last point we examine another theory of explanation which lacks such an examination of dependency structures. In this section we will briefly discuss Gardenfors[16] analysis of explanation and show that its underlying motivation is effectively compatible with our own. We briefly recall the nature of AGM belief revision.[17]. The idea is to represent beliefs and knowledge syntactically, and then define operations over our knowledge and beliefs. As discussed earlier assume T is a theory, and $\vdash$ is our consequence relation, and Cl(T) is the closure of T under logical consequence. Then we define an update operation ($\circ$) axiomatically. Assume T is a set of sentences in the language $\mathcal{L}$, a belief set.

- (1) $\mathrm{T}\circ\phi$ is a belief set and (2) $\phi \in \mathrm{T}\circ\phi$.
- (3) $\mathrm{T}\circ\phi \subseteq \mathrm{Cl}(\mathrm{T}\cup\{\phi\})$ and (4) if $\neg\phi \notin \mathrm{T}$, then $\mathrm{Cl}(\mathrm{T}\cup\{\phi\})\subseteq \mathrm{T}\circ\phi$.
- (5) $\mathrm{T}\circ\phi = \mathrm{Cl}(\bot)$ iff $\vdash_{\mathcal{L}} \neg\phi$ and (6) If $\vdash_{\mathcal{L}} \phi \leftrightarrow \psi$, then $\mathrm{T}\circ\phi = \mathrm{T}\circ\psi$.
- (7) $\mathrm{T}\circ(\phi \wedge \psi) \subseteq \mathrm{Cl}(\mathrm{T}\circ\phi \cup \{\psi\})$ and (8) If $\neg\psi \notin \mathrm{T}\circ\phi$, then $\mathrm{Cl}(\mathrm{T}\circ\phi \cup \{\psi\}) \subseteq \mathrm{T}\circ\ (\phi \wedge \psi)$

Most of these axioms are self evident and the ($\ominus$) retraction axioms are defined analogously[18]. The most interesting cases are (7) and (8) since they relate to the idea of iterated belief revision, and how the sequential revision requires consistency checks. In particular (7) is related to (3) and (8) to (4). We can allow that there is a conditional probabilistic model of any given theory, where we define $\mathrm{T} = \{\psi \mid \sigma(V(\psi)) = 1\}$, $\mathrm{T}\circ\phi = \{\psi \mid \sigma(\mathrm{V}(\psi) \mid V(\phi)) = 1\}$ and similarly, if $\mathrm{V}(\phi) = \emptyset$, then we define $\mathrm{T}\circ\phi = \mathrm{Cl}(\bot)$. These notes are a far from adequate representation of the complexities of belief revision theory. However, this is adequate preamble for introducing Gardenfors definition of explanation. The core idea of Gardenfors is that an explanation is to be sought by initially withholding belief in the *explanandum* and trying to determine what if anything would induce belief in the *explanandum*. Notation: Let K^{-}_{ϕ} (B^{-}_{ϕ}) denote that ϕ has been contracted from the set of known (believed) sentences, and K^{+}_{ϕ} (B^{+}_{ϕ}) say that the set of known (believed) sentences has been expanded to include ϕ. Similarly $\mathrm{K}^{\circ}_{\phi}$ ($\mathrm{B}^{\circ}_{\phi}$) denote that our knowledges and beliefs have respectively been revised with ϕ. Both K and B are here defined in terms of a probability function, where K is equated with certainty, and B covers any degree of belief of belief less than certain.

Definition (Explanation) An explanation of a singular sentence ϕ relative to a state of knowledge K (where $\phi \in$K) consists of (i) a conjunction P of a finite set of probability sentences and (ii) a conjunction C of a finite set of singular sentences that satisfy the requirements that (iii) $\mathrm{B}^{-}_{\phi}(\phi \mid P \cap C) > \mathrm{B}^{-}_{\phi}(\phi)$ where B^{-}_{ϕ} is in the state K^{-}_{ϕ}, and (iv) $\mathrm{B}(\mathrm{P}\cap\mathrm{C}) < 1$ (that is, $\mathrm{P}\cap\mathrm{C} \notin$ K).[19]

Crucially here, the idea is that we seek a probabilistic dependence relation to determine that ϕ will be occurent on the basis of some precise conjunction P∩C, which we believe to be true whether or not ϕ occurs. The idealisation here is that our beliefs are already truth-tracking, i.e. any update of T with ϕ can be motivated by some beliefs already present in T. This idea underlies our theory of explanation as well, but our focus is more explicitly on the nature of the dependency we hope to discover. The main problem with Gardenfors approach relates to the problem which was our motivation. Stated as follows:

> The source of the problem has long been identified by philosophers of science: ‘some regularities have explanatory power, while others constitute precisely the kinds of natural phenomena that demand explanation’. The regularities captured by the probability

[16] cf. [35]pg167-180

[17] This discussion draws from [93], [45] and of course [2]

[18] To be found in [93]pg45 -47

[19] [35]pg178.

> sentences in Gardenfors definition will often have no explanatory value even if they make the explanandum completely unsurprising[20] . Without further restriction on the explanans, the account will always be vulnerable to such counterexamples.[21]

In effect this means that the intellectual satisfaction we expect of an explanation requires that an agent who given an answer to a why-ϕ question must be prepared to identify such answer as being properly relevant to the question asked, i.e. appearing in the partition determined by Q_{ϕ}. These observations suggest exactly the considerations we have been attempting to motivate.[22] Pearl argues that these considerations are apt to refute the "associational criterion" of when a statistical correlation is non-confounding, thereby motivating the requirement that direct a direct dependency criterion is required to define the notion of a non-confounding correlation. In particular we can show that there is always a way in which the observation of a particular "association" or correlation is never alone sufficient to ensure that we have a non-confounding result.[23]

6.2.6 Gettier Again: Knowledge as Convergence under Conflict Resolution

Gettier problems seem unavoidable, and hence they motivate the view that knowledge must be more than justified true belief. A number of attempts have been made to specify what fourth condition would save knowledge from remaining unattainable. To my mind the best suggestion has its roots in Pierce's definition of truth as the limit of inquiry[24] and Lehrer's notion of knowledge as those truth claims which survive the scrutiny of the scientific community.[25]. Formal attempts to incorporate this notion can be seen in the work of Baltag and Smets[26], where they prove some convergence theorems for update operations such as ($\Uparrow$), defined above. We shall close our discussion of explanation with the suggestion that similar results will hold for updates with the explanation operation ($[\xi^{\psi}_{?}{:}\phi]$) in a particular community if each candidate explanation can be defended immediately after deployment in such a way that every why-question in a community can be adequately resolved one way or another after the appropriate justification game has played out. This kind of justification game is played out until some faction wins, or every faction loses i.e. is forced to accept an inconsistent information state. Grant that the justification procedures are shared by all participants, then all disagreement amongst factions can be tracked back to a disagreement at the level of their dependency models. For a faction to win this game they could win every round of dispute, or just not lose a round which forces a contradiction with their dependency information until the groups questions were exhausted. But we could also allow for a kind of consolation prize. The faction that survived the longest series of explanatory update and conflict resolution without engendering inconsistency could be declared interim victor. This is a viable criterion for model preference. The details are not trivial[27], but consider how this idea could be used to address Gettier's argument.

The Gettier argument works by introducing novel, typically absurd, considerations into the contrast class of answers for any given why-question. Our suggestion is that if we can specify a set of possible candidate answers for each why-question based on considerations of our implicit dependency theories, then we exclude the possible introductions of underdetermination possibilities. Furthermore, if the community mandated defence procedures for any answer in our justification

[20] Think of the correlation between a functional thermometer and a heatwave.

[21] [71]pg142.

[22] This point is also made by Pearl and Halpern in [47]

[23] He also shows that the "associational criterion" won't even meet the necessary condition for ensuring non-confounding effects i.e, that no association ever results in a confounding result when it meets the appropriate definition cf. [73]pg182-188

[24] See [74]pg268

[25] See [61]

[26] In particular see [10]

[27] In particular, we would need to further flesh out the notion of a justification game. In this we can draw on the work of deBruin; for instance see [24].

game rely upon dependency information then no grue-type answer will be defensible if our dependency information excludes such underdetermination possibilities. This, we would hope, is sufficient to define a species of knowledge which is attainable in practice. In other words, this is proposal to define knowledge as justified true belief in the limit of inquiry - where the limit of inquiry is thought of as the a bound determined by the number of questions entertained in a group. However, all these possible avenues of exploration are only available to us if the underlying notion of dependency is sound.

Identifying Dependence

We have suggested that the best way to pursue genuine explanations is to look at the dependency structures which underlie our model (view) of the world. The natural objection to this line of thought is to question the existence of such dependency structures. Call this the debate over structural realism. We now elaborate a defence of structural realism against van Fraassen's alternative of constructive empiricism.

6.3 Structural Realism

Suppose you're in a desert, trudging through the sand endlessly hoping for an oasis over the next horizon. You know from past experiences that a mirage is a dangerous thing but upon sighting a playground, with swings, slides and monkey bars, you head for the structure. The structural realist believes that the playground exists, and once we arrive we can climb the monkey bars. In the same way we claim that you should be a structural realist about the dependency structures of the world. In general, they are much more useful than monkey bars.

Structural Realism There exists a dependency structure of the world, that we can come to know. Our best theories of such structures track a variety of dependence relations by which the world is stratified.[28]

Opposed to this view is the conventionalist or constructivist view of dependency relations. The constructivist claims that the image of the playground is a mirage cobbled together by the bizarre propensities of a desperate mind. Similarly, where the realist talks of the dependency structure of the world, the constructivist speaks of the imagined, or definitional structure of the "world". Perhaps the most extreme example of a constructivist is Nelson Goodman whose thesis modestly begins: "[w]ithout presuming to instruct the gods or other worldmakers...I want to illustrate and comment on some of the processes that go into worldmaking".[29] No mere playground suffices for Goodman's imagination. The very world in which we each live is, for him, a construct, a fanciful fabrication. Importantly, this attitude is primarily based on his grue-problem and the associated underdetermination results. A more contemporary picture of the constructivist idea is articulated by David Chalmers in his recent book *Constructing the World*, wherein he advocates for a scrutability thesis - which states that all truths about the world can be derived *a priori* from the appropriate set of foundational truths.[30] Chalmers is technically agnostic about the reality of the structural relations which underlie his scrutability thesis, but Goodman is insistent that such structures are always and only figments of an over active imagination. We will contest this idea, although not directly.

Constructive Structuralism All structures which we come to know are, by and large, a product of our creative imagination motivated by more than what we can actually observe.

[28] This terms is a little overloaded since there is another position of the same name discussed for instance by Frigg and Votsis in [34]. Our position of structural realism is related, and even supported to some extent by their arguments, but it is worth noting that there is a subtle difference in that the articulation of the position.

[29] [41]pg7

[30] [20]pg430

The real point of dispute between the realist and the constructivist involves a concern for an economy of theory development, and the limitations of inference given evidence. When asked how the realist assumption helps us navigate the world, the constructivist argues that the assumption is of no help. As such the burden of proof appears to be upon the constructivist. It's not. We shall first show that van Fraassen has developed a kind of reasonable constructive empiricism regarding the methods and conclusions of natural science. Given that he appears to do so without the assumption of structural realism, this places the burden upon us to show that either his position is (a) false (b) inadequate or (c) incoherent. We shall do all three.

6.4 Constructive Empiricism

Technically van Fraassen's target is scientific realism, which can be thought of as an extension of structural realism depending on how broad we construe the domain of science. His arguments against scientific realism are compelling precisely because he gives a kind of weak characterisation of the scientific enterprise. Such an enterprise is often (he argues) assumed falsely to be accurately characterised as follows[31]:

> Science aims to give us, in its theories, a literally true story of what the world is like; and acceptance of a scientific theory involves the belief that it is true.

On such a characterisation we see science as a fallible enterprise aiming to achieve the truth, where a minimal condition for the truth of a theory is that its acceptance induces belief in that theory. We take this to be a fair characterisation. A consequence of this view is that science can be seen to accept the existence of unobservable entities, or relations where such things feature in the claims of an accepted scientific theory, and then by definition we are expected to believe in the existence of these entities and relations. Van Fraassen objects to the latter conclusion; belief, he thinks, is not necessitated by acceptance of the theory. He is motivated by broadly empiricist leanings, whereby all those things unobservable in nature are open to disbelief.[32] In a very strong sense, seeing is believing for the constructive empiricist. As such, the adequacy of a scientific theory is to be tested only on the action and behaviour of the observable entities that fall under the remit of the theory. However it's not clear as to whether an unobservable element of one theory will forever remain unobservable given the march of technology. This prompts the following (very general) constructivist account of the scientific enterprise:

> Science aims to give us theories which are empirically adequate; and acceptance of a theory involves as belief only that it is empirically adequate.[33]

This latter constraint on belief allows us to incorporate a theory within our beliefs in so far as its acceptance lends itself to some practical application - however this is construed. In such a way van Fraassen can argue for the adoption of a particular theory of causation for its predictive benefits, and so there seems no bar for him to adopt the best scientific theories whatever their initial motivation so long as he can find some ultimate use for the theories. In effect scientific theories are accepted to be fundamentally underdetermined. However there seems to be a continuum here in terms of the degree to which a theory is useful or practical. So with such a pragmatic motivation we might worry that the underdetermination of a theory's truth can be extended to the question over whether a theory is empirically adequate? Surely grue-type cases emerge here too, if there no restraint on rational acceptance other than empirical adequacy and empirical adequacy is purely a pragmatic measure? The natural objection here is to insist that empirical adequacy is a measure determined by the objective correctness of a theory with respect to the actions of its observable parameters.

[31] [32]pg8

[32] I'm showing my hand here. Strictly speaking the constructivist need not tolerate the suggestion that anything unobservable, properly speaking, exists.

[33] [32]pg12

> To present a theory is to specify a family of structures, its *models*; and secondly, to specify certain parts of those models (the *empirical substructures*) as candidates for the direct representation of observable phenomena.[34]

A theory is defined as *empirically adequate* just when some model of the theory can be put in isomorphism with the observable facts, i.e. when the empirical substructures of the model reflect all the observable facts. Note, it is not true to say that a theory is empirically adequate just when all empirical substructures of said model can be put into an isomorphism with some of the facts. Putting all this together, we see that theory acceptance is to be mandated by empirical adequacy on pain of the obvious underdetermination problems for liberal pragmatism. The more observable facts reflected in the theory the better the theory. With this in mind, we now turn to the arguments against scientific realism.

IBE does not entail Realism

A cheap construal of abductive reasoning might appear to offer the suggestion that the success of scientific theorising is apt to induce belief in the existence of the entities (observable or unobservable) postulated in the theory. However, this is not totally accurate. We can, with van Fraassen, be voluntarists about beliefs - whereby IBE may appear in the canons of rational inference along with the probabilistic axioms but in no way does the acceptance of these axioms necessitate their usage, or belief in their consequences. So for instance we might see that our mouse trap has been activated, the cheese has disappeared and remain adamant that no mouse exists behind the wall even though IBE would suggest the contrary. This argument is sound. However, it is uncompelling since presumably we still seek to answer the existence question which prompted the argument. The empirical fact that people can be seen to ignore inference to the best explanation, does nothing to undermine the realist thesis. The argument does show that realism is not the only contender in this debate...we could also be inexplicably agnostic, or suggest reasons to doubt the existence of mice.

A better argument against the realist position is deployed by van Fraassen to indicate that realism necessitates an explanatory regress. The idea is used to undermine the power of IBE. He writes...

> ...even if we were to grant the correctness (or worthiness) of the rule of inference to the best explanation, the realist needs some further premise for his argument [in each case of IBE]... So the realist will need his special extra premiss that every universal regularity in nature needs an explanation, before the rule will make realists of us all.[35]

This argument would seem to seriously problematise the idea that realism follows by appeal to IBE. But it is not at all evident that the realist's search for explanation is so fevered. Indeed we might think that belief in the existence of a realist theory which appears to explain some phenomena is motivated precisely because it explains said phenomena (without being *ad hoc*), not because we feel some inordinate need to find or introduce novel explanatory theories at each instance. Van Fraassen's argument targets the philosophical use of IBE to motivate realism wholesale, but realism is always properly motivated piecemeal. Nevertheless van Fraassen's argument is correct as far as it goes, which is not very far.[36]

[34] [32]pg64

[35] [32]pg21

[36] More forcefully, we could make the argument that the structure of a theory is all we can come to know, either in the sense that it is all we can come to cogently communicate or perceive. For details on this species of argument see [34]

Theory Laden Observation

The insistence that all scientific claims are theory laden is often taken to be a point against realism of any stripe. But it's easy to concede that there is some truth to the observation without troubling the realist thesis. Although our language might well incorporate various theoretical predicates, most such predicates have been well motivated before their widespread inclusion in our scientific theories. As for the notion that the concept of observability is theory laden - van Fraassen accepts the charge that even if the limits of observability are determined by community consensus, then so be it, "it is, on the face of it, not irrational to commit oneself only to a search for theories that are empirically adequate, ones whose models fit the observable phenomena, while recognising that what counts as observable phenomenon is a function of what the epistemic community [insists]".[37] But by similar reasoning it is not irrational to commit ourselves to a search for causal (supervenience, grounding...etc) explanations while recognising that the criteria of adequacy on such searches is fundamentally determined by community consensus. In short, the charge that our account of dependency based explanation is riddled with theory-laden vocabulary cuts both ways, and is not obviously worrying in either case. A cogent objection to structural realism must further demonstrate that our theory incorporates a *harmful* theoretical hangover from our model of e.g. causal dependency, in which case we enter into debate over how to assess the adequacy or harmful nature of our assumptions. Such a discussion is nothing more than a debate over model preference, so at worst such a discussion only improves our model.

Structural Realism is not the only route to Success

You might want to challenge the constructivist on the basis that they seem to collapse the line between useful and correct theories. While the realist can explain dependency relations between particular events by appeal to deep structure, the constructivist appears prone to denying that there is an unobservable factual structure of dependence. Instead he allows for only definitional or correlative connections based on a history of observation. Again, this realist argument falters because it begs the question we are trying to determine. If truth is a mark of success then the ability to account for presumed true relations of dependence will always make the realist theory more successful than the constructivist theory. By equating explanatory power with the ability to identify unobservable structure, then we deny that there is any explanatory power to be derived from statistical information of *observed* correlation between events. This is false. The success of a theory may be measured in practical terms without question-begging appeal to supposed truth-approximation of our hypothetical deep structure.

In the same vein we might think that since scientific explanations of the properties of objects are often based on the hypothesised unobservable microstructures, the constructivist would be at a loss to develop analogous but strictly empirical explanations for the same phenomena. However, van Fraassen can argue that a difference in microstructures will have observable consequences, hence given enough time and data the constructivist will be able to infer from the statistical correlation some serviceable property of the entity in question to facilitate predictions. This property can be fed into all predictive (or otherwise scientific) uses of this entity in our theory so that wherever the realist supplies an explanation so too the may the constructivist given enough effort. Worse news for the realist is that the microstructure hypothesis allows for the development of an indeterminate number of alternative theories with subtly different assumptions about the microstructure of the world. This effectively makes any choice of microstructure a result of some "hidden variable" which determines the choice. Such a "hidden variable" must be made explicit if it is to be rendered reasonable. By making such assumptions explicit we can see science as a history of competing narratives where each narrative "suggests new statements of observable regularities and...corrects old ones"[38]

[37] [32]pg19

[38] [32]pg34

The Ultimate Argument does not ensure Realism

The final argument for realism is dubbed the ultimate argument by van Fraassen who clearly wants to draw attention to the misnomer. The idea, initially propounded by Putnam is to argue that the success of our theories, the accuracy of their terms, their predictive power, etc are all rendered miraculous if our theories are in fact false but simply empirically adequate. The challenge is to say that there must be an empirically adequate reason for why we converge on empirically adequate theories. Putnam suggest that the actual truth of the theory is the best candidate reason. Although, given the picture of the scientific enterprise which van Fraassen has developed you might predict his answer:

> The Darwinist says: Do not ask why the *mouse* runs from its enemy. Species which did not cope with their natural enemies no longer exist. That is why there are only ones who do. In the same way, I claim that the success of current scientific theories is no miracle... For any scientific theory is born into a life of fierce competition, a jungle red in tooth and claw. Only the successful theories survive - the ones which *in fact* latched on to the actual regularities in nature.[39]

As such, he argues we should expect that our arrival at the acceptance of an empirically adequate theory is well motivated and not accidental. So construed we have a plausible constructivist view of the scientific enterprise. Hence allowing van Fraassen's construal of scientific activity we come to the primary empiricist conclusion: "the assertion of empirical adequacy is a great deal weaker than the assertion of truth, and the restraint to acceptance delivers us from metaphysics" Hence it appears that our efforts in the previous chapter were improperly motivated.

6.5 Pragmatics of Explanation

With these considerations in mind we can state the first and probably most troubling objection to our project.

The Adequacy Objection: An explanation need not be true to be accepted

The worry here is that if the truth is irrelevant to explanatory power then our project to develop an account of explanation on the basis of a truth-tracking relation over actual dependency structures seems misguided, or at least needlessly complicated.

6.5.1 Explanation as Answer

One might understand the objection as the claim that our theory is overly complicated and should be rejected on the basis of Ockham's razor. To make this assessment we would have to see to what degree the pragmatic account of explanation is indeed simpler, if at all. We shall now briefly elaborate van Fraassen's theory of explanation.

The constructivists idea is that, to"say that a theory explains some fact or other is to assert a relationship between this theory and that fact, which is independent of the question whether the real world, as a whole, fits that theory." Such an assertion explicitly elides any commitment to the truth of the theory. The view is supported by some choice quotes from Darwin, Huygens and Newton, luminaries all. This seems to force van Fraassen to the conclusion that the "explanation-relation is visible before we believe the theory is true".[40] This claim is somewhat puzzling unless we accept the idea that there is a semantic relationship between a particular theory and claim. However, if such a semantic relation exists is it part of a van Fraassen's model of explanation? If

[39] [32]pg40

[40] [32]pg98

so, what is the status of this model? Its existence seems to force van Fraassen's commitment of abstracta, and hence unobservable relations? But lets accept the claim for the moment at face value.

If the claim were true, then we have developed a theory which effectively stipulates that one should always search for true explanatory relations. I would profess a great deal of surprise if any of van Fraassen's quoted luminaries objected to this normative ideal. So then the process of developing models of dependence relations on which to base different species of explanation, becomes a process of attempting to sort the simply adequate models from the true model. The constructivist is not yet in a position to remove truth from the set of desired norms. So our task is just different from the more modest constructivist task.

Digging deeper we see van Fraassen develops a theory of explanation where explanations are all answers to "why-questions". Effectively this is similar to the idea that an explanation is a solution to an abductive problem where we take a "why-question" Q to be determined by three parameters: (1) the topic T_k (2) the contrast class X of potential answers (solutions) (3) the relevance relation R.[41] The idea is that the topic T_k will provide us with a set of presuppositions to be incorporated in viable candidate models, while the contrast class is akin to the set $\mathrm{X} = \{\gamma, \gamma_1 ... \gamma_n\}$ of abductive solutions to an underdetermination problem. Finally, the relevance relation R is an unanalysed primitive. Finding an answer for any question is therefore relative to a particular context which validates the presuppositions of the topic, and ensures that each answer to the question is **fully explanatory** given the context.[42] The definition is as follows:

Definition (Direct answers) Γ is a *direct answer* to a question $\mathrm{Q} = <\mathrm{T}_k, \mathrm{X}, \mathrm{R}>$ exactly if there is some proposition ϕ such that ϕ bears the relation R to $<\mathrm{T}_k, \mathrm{X}>$ and Γ is the conjunction of ϕ and T_k and the negation of the all parameters in the topics T_i such that $\mathrm{i} \neq \mathrm{k}$.[43]

This approach prompts van Fraassen's reflection that traditional construals of explanation go wrong in so far as they fail to appropriately factor for the details of context. He wishes to conclude that there is no one principled relation called scientific explanation, since a scientific explanation is simply a direct answer to a question which makes use of scientifically derived information so as to determine which is the appropriate answer to the stated question. We are in agreement on this issue with van Fraassen, but we simply suggest that the manner in which we determine the appropriate answer can be explicitly related to our theory of dependence relations and the logic of such structures. This is more informative than broad appeals to an intransparent notion of relevance. Furthermore we insist on the extra normative condition that an explanatory answer ought to command assent in a given community where the appropriate contextual information is accepted. In short, we wish to say that an explanation is a true answer to a question which satisfies all in ear shot.

Our theory differs from van Fraassen's because we do not insist that an abductive solution (or answer) must be **fully explanatory**[44] and we do not rely on an unanalysed notion of relevance. There are non-predictive explanations which do not rule out all other candidate explanations, and hence prompt subtly different changes in the doxastic and epistemic states of the agents in receipt of such explanatory information. Our approach allows us to incorporate explanations of differing species and strengths on the assumption that such variety is required to make sense of the behaviour of rational agents. For both theories the crucial component relates to how we should evaluate particular candidate explanations - van Fraassen cites Simpson's paradox as a representative case

[41] Note that this is a somewhat distinct notion of a question-as-partition than we have previously discussed.

[42] As defined in chapter four.

[43] [32]pg144

[44] To be fair van Fraassen is aware of the strength of this condition, but seeks to motivate it by appeal to a Gricean ideal of effective communication cf. [32]pg144. This move is fine as far as it goes but it hints a normative constraint not necessarily operative in every domain in which explanation is required.

in so much as there is no obvious way in which to choose amongst the explanatory candidates so as to determine which is "telling, good, or better" without directly incorporating the "relevant" probabilistic assumptions.[45] As we have already shown[46] Simpson's paradox dissolves under the proper construal where we directly incorporate structural causal information.

Response to the Adequacy Objection The apparent extra complexity of our theory as developed in this thesis is not needless, indeed it fills some of the gaps in van Fraassen's model. Hence it cannot, and should not, be dismissed by an application of Ockham's razor. Better, it is not *ad hoc* since structural theories are not without broader motivation contrary to more obviously question begging stipulations of relevance.

Arguably our theory is normative where van Fraassen aims to be broadly descriptive of explanation as practised. To motivate this normativity we shall have to show that the search for truth (rather than empirical adequacy) is a hallmark of the scientific (or more generally rational) enterprise - we will make this case below.

6.5.2 Evaluating Answers

In a way van Fraassen's set up appears much more general than ours due to his ability to factor for the dynamic features of relevance given context. As such we might expect an objection to our project based on its limited utility.

The Scope Objection Only limited explanations can be developed on the basis of dependency structures.

The worry here is that our theory is too inflexible to be an adequate theory of general explanation. The thought is that at best we have isolated a subspecies of explanation.

6.5.3 Counterfactual dependence and Relevance

The inherent flexibility of van Fraassen's model stems from the fact that he can define conditions of relevance so as to preserve a relation of counterfactual dependence between the *explananans* and *explanandum.* This is in part possible because van Fraassen seems to doubt that counterfactual conditional claims are properly speaking truth-apt. They have nothing but an explanatory role in theory construction - they highlight accepted moves of inference present in our active theory. He claims, "we must conclude that there is nothing in science itself - nothing in the objective description of nature that science purports to give us - that corresponds to these counterfactual conditionals."[47] This idea stems ultimately from his preference for observable facts, and the idea that no counterfactual claim is ever clearly testable. Worse their evaluation, he claims, is ultimately context sensitive and no such claims ought to appear in the scientific record. As such, van Fraassen allows that we may gerrymander the counterfactual claims in our theory such that they respect our intutive notion of relevance and support inferences to prediction that can in their turn be observed to be true. But such heuristics only play a pragmatic role in discussing scientific facts, they are not themselves truth-apt in any objective sense.

The flexibility of van Fraassen's theory is indeed a virtue, but it is not a virtue which necessarily escapes our own theory. The ability to inaugurate counterfactual conditionals on the basis of a presumed dependence is something we which our theory attempts to motivate. The only difference is that we take seriously the idea that counterfactual conditional claims have objective truth conditions which may only be properly assessed if we have broader theory of the relation (i.e.

[45] [32]pg146

[46] In the previous chapter.

[47] [32]pg118

causality, supervenience, grounding etc) on which the counterfactual considerations are based. Worse, if we do not explicitly restrain the notion of relevance, van Fraassen's theory of explanation can be trivialised whenever we gerrymander the relevance constraints. For instance, we can make ϕ the only relevant answer to the why-ϕ question by suitably weighting the context.[48]

Response to the Scope Objection Our theory can be rendered as flexible as van Fraassen's if we presume that for every counterfactual conditional he admits as a pragmatic heuristic there is an underlying relation of dependence, which we could find and motivate independently in the scientific enterprise. The nature of this dependency structure might be subtle, and the manner in which we draw out the connection between the *explanans* and *explanandum* need not be obvious. The point is that if we find such relations we take ourselves to have circumscribed some of the conditions which make the latter relevant to the former. Such connections need not be *ad hoc*, so long as there is a broader motivation for the theory in which they are found.

The scope objection draws attention, not to a deficit of our theory, but to a crucial feature of van Fraassen's theory. His stance on the reality of modal and counterfactual claims of any stripe is such that it requires a defence not forthcoming. We shall examine an argument below which uses this fact to devastate van Fraassen's project.

6.5.4 The Quantum Counterexample

Our procedure seems to relate explanation to the observation of broadly deterministic procedures. If we can find an irreducibly indeterministic system than our model of explanation is undermined.

The Quantum Objection Explanation in quantum mechanics requires that the underlying system is entirely indeterministic.

This is true, but the fact imposes a problem on our model of explanation only if a focus on dependency relations rule out indeterminism. They don't. As we have been at pains to demonstrate the dependency relations between events can be represented purely in terms of probability and/or deterministic links. For our theory to accommodate explanations in quantum mechanics we need only drop the pretense that there is an underlying mechanistic function which relates the relevant quantum events. As Batterman argues it crucial to avail of limit arguments. Replacing a mechanistic function with a statistical law that the probability of certain events converge to 1 in the limit when appropriately conditioned.[49]

Response to the Quantum objection There are far more controversies over the interpretation of probability in quantum mechanics than we have space to comment on, however the main point is that nothing about our theory of explanation via dependency structures rules out a purely probabilistic model of such conditional dependency relations amongst quantum phenomena.

If we wish to flesh out these dependencies as observations made over a "good family of ideal experiments" as van Fraassen[50] recommends, we can. It is an open question over whether we should.

6.6 Structural Realism: An Offensive Defence

We now wish to consider some objections to van Fraassen's project. These should be seen as an indirect defence of our own position. Profitably these arguments can be seen as an attempt to roll back the significance of van Fraassen's initial objections to scientific realism.

[48] Similar observations are made by Kitcher and Salmon in [55]

[49] [12]

[50] [32]pg190-194.

6.6.1 Evolutionary Epistemology and The Ultimate Argument are consistent

We wish to make two points here: (1) Nothing about the evolutionary reasoning that van Fraassen uses to motivate a belief in the empirical adequacy of our accepted theories, is inconsistent with the Putnam's brave conjecture, (2) the evolutionary reasoning is insufficient to assure convergence on an empirically adequate theory, rather the best we can expect to achieve is a winnowing of decreasingly less adequate empirical theories.

The law of the jungle ensures that the least adequate theories are rejected, but it does nothing to ensure that the actually adequate theories are the most prevalent. Although (2) is sort of conceptually obvious, we can cite an argument to the same effect. Darrell Rowbottom applies Price's equation for transmission and selection of traits in a given a population to assess the viability of theories under pressure of test, where the inherited trait is taken to be "empirical adequacy". He shows it fails to converge on actual empirical adequacy.[51] The argument informally, is something like this: even on the assumption that we are in receipt of accurate observational information there is a continuum of exactitude which we can apply to predicting the action of observables. Think of the observation that the kettle boiled. We might say it boiled "now", and expect to be understood, but on inspection we can zero-in on the exact time only by making an infinity of adjustments of our initial estimate over decreasing intervals of time. Each observation being more "empirically adequate" than the last. As such, we falsify an infinity of empirically inadequate theories without ever arriving at a properly speaking adequate theory.[52] The kettle example tries to demonstrate that there is a continuum of exactitude for which no preference relation actually holds. So despite holding to a preference for empirical adequacy we can winnow an infinity of theories, and expect to winnow an infinity more, without ever reaching the conclusion that this one theory is properly empirically adequate for all cases. This is made worse, if we think of exactitude not with respect to a temporal parameter, but instead as a locational parameter with respect to distance from a particular landmark. In effect this is an underdetermination argument since I need various sci-fi style augmentations of our observational capacity to keep extending the limit in perpetuity. If we get bored we might call our latest theory empirically adequate, but this is not because we have met the definition of perfect prediction (description) of observables. If we aim only at truth we do not require or even motivate an incessant search for inane exactitude, nor are we forced into theory change because some unimportant fact is technically speaking better accounted for by the novel theory. Criteria for adequacy on a theory is simple, if truth is a constraint.

The argument here might seem a little flippant, but the important point it raises is that there is a tension between defining empirical adequacy in terms of observable action where there are non-obvious, and not yet definitively reached limits on what counts as observable while we want to converge on an empirically adequate theory. The convergence is impossible if empirical adequacy is defined by observable parameters under a nebulous notion of exactitude, but if we allow that the limits of observability are fixed by definition then convergence to an arbitrary limit is easily achieved. The latter is likely to be the course pursued by van Fraassen but this strikes me as a strange criterion of adequacy on an empiricist picture of scientific adequacy. How is he to avoid the objection that adequacy of a scientific theory defined by a local criterion of observational adequacy does anything but beg the question against reasonable alternatives in an entirely unprincipled manner? He needs to be able to state that as an empirical fact there is a limit of our observational ability for the convergence argument to work, but this is not an *empirically established* fact.

6.6.2 Microphysics does not entail Constructive Empiricism

The idea here is that we take issue with van Fraassen's disdain for theories with "hidden variables". The role of hidden variable theories (HVT) in scientific practice is not purely detrimental to the

[51] [78]

[52] This is somewhat akin to Zeno's paradox.

enterprise. As such much of the motivation for constructive empiricism collapses. In this argument we follow Silvio seno Chibeni[53] who presents a rational reconstruction of the reasoning of Einstein, Podolsky and Rosen regarding the interaction of two spatially separated quantum particles. The problem, in short, seems to be that we can observe a qualitative correlation between the states of the two particles for which we cannot offer an explanation if we take quantum mechanics as offering a complete description of these particles. So we are left with the option that either quantum mechanics is incomplete or there is some global relation of cause and effect which we cannot discern at a local level. The hypothetical global mechanism is, for methodological reasons, disdained. Hence quantum mechanics as it is currently stated should be thought of as incomplete. This prompts the hypothesis of hidden variables.

Discomfort with HVTs is usually motivated by showing that (depending on the nature of the postulated hidden variable) we can (i) derive an inconsistency with quantum mechanics as it stands, or better (ii) the hidden variable assumption is usually supposed to be a total function defining the absolute behaviour of the two particles, and it was shown by experimental research in the 1960s that no such uniform (or global) function accurately reflects the (local) empirical variance in the attributes of the relevant particles. Hence HVTs seem to be contrary to both the methods of quantum mechanics and empirical facts. Van Fraassen takes these observations to indicate that HVTs are guilty of a question begging search for "correct" explanations and we should no longer search for explanations when instead we can take the EPR-correlations as brute (but useful) fact. However, on a more general level David Bohm has developed a consistent and somewhat scientifically fruitful theory which incorporates global hidden variables. So we might expect van Fraassen to provide good arguments against such a HVT?

Chibeni argues that van Fraassen appears to be motivated by an unfounded worry that such theories can come into conflict with relativity theory. But even assuming they conflict, this is no bar on further developing quantum theory with hidden variables - whether the theory is ultimately more or less profitable can only be decided in retrospect. It is fallacy to assume that there is no merit to postulates simply because their assumption involves a questioning of orthodoxy.[54] Something akin to this line of objection is used to counter each (of six) types of objection van Fraassen has against HVTs. The result is that van Fraassen's position is seen to be increasingly dogmatic and his insistence that we should not seek for explanations for these EPR correlations is entirely uncompelling.[55] Chibeni concludes that van Fraassen's arguments "make the prospects of explaining the EPR-Bell correlations look darker than they actually are... the results do, impose severe constraints on the most natural explanations for these quantum correlations, but, provided we are willing to pay the appropriate price, the classical scientific ideal of explaining the natural 'mysteries can be retained as a valid and stimulating intellectual challenge"[56]

6.6.3 Constructive Empiricism requires Abstracta

Gideon Rosen spends significant effort mulling over the proper interpretation of van Fraassen's theory of constructive empiricism. Dismissing the idea that it is either straightforwardly descriptive or normative he comes to the view that van Fraassen elaborates an *as if* story about nature of the scientific enterprise. He does so on the basis of what can be construed as stylistic preference - not comfortable with the idea that scientists can speak (philosophically) for themselves, van Fraassen puts forward constructive empiricism as a type of narrative describing a minimalistic interpretation of the activity of scientists.

[53] [22]

[54] [22]pg59

[55] Alan Chalmers argues similarly that van Fraassen's characterisation of the significance of Perrin's experiments on Brownian motion is entirely misleading. Contrary to van Fraassen's conclusion Chalmers makes the case for being realists about the existence of atoms on the basis of Perrin's arguments in [19].

[56] [22]pg63

> The fictionalist reading of constructive empiricism does justice to both the descriptive language of the texts and their probable falsity if taken literally. Much of what van Fraassen says is profitably seen as an effort to make good the claim...that science 'makes sense' when interpreted as the search for empirically adequate theories. It also discloses a pleasing symmetry in van Fraassen's larger view. Just as science is portrayed as an *as if* story about nature, so the authors own remarks about science are to be taken as fictional assertions expressing a commitment to this idea as an adequate *as if* story about the intentions of science.[57]

The important point here is that we can come to see constructive empiricism as a kind of normative claim about how to practice science, if we accept the empirical adequacy of his description of science as practised. So by Ockham's razor no more extravagant theories should be accepted, or indeed believed. Given this background Rosen suggests that van Fraassen, himself, fails to adopt the normative self-constraint that all scientific results can be treated as if they were derived from observable experience. This failure is primarily due to manner in which van Fraassen checks the adequacy of scientific theories. Namely, he suggest that we draw up semantic models of the state space and test our models for accuracy against the behaviour of the observable elements featured. However, as Rosen is quick to suggest...

> Any account of theory acceptance as the belief that one's theory possess some good-making feature Q will imply that to accept a theory is to take commitment to the existence of theories, and so to abstract [unobservable] objects.[58]

Given that we have enough acquaintance with our models to determine their accuracy, we should not be surprised to hear statements such as "$\exists$(M, w) such that M, w $\models$ T$_i$"... and so on for all amended theories, where w is the actual world. In short, we quantify over models in the elaboration of constructive empiricism. But no such information about the existence of models (or their properties) was derived directly from observable experience. Hence, van Fraassen's meta-story about the course of science is itself not adequate by its own standard. So there is no normative force to apply Ockham's razor and believe in constructive empiricism. Given that the existence of theories is so crucial to van Fraassen's account of the scientific enterprise, renouncing belief in the abstracta is tantamount to renouncing belief in constructive empiricism altogether. In short we have an *as if* story about the course of scientific practice which fails its own test. Does he then wish to strongly distinguish between the activities of science and philosophy? How could the latter be worthwhile if it differs from science, given the presumption that all information has an empirical source? Is this presumption not the hallmark of empiricism? What role (if any) should van Fraassen's account have in our considerations about scientific reasoning?

6.6.4 Constructive Empiricism is inconsistent or perverse.

In this section we recount an argument indebted to James Ladyman which purports to reveal an inconsistency at the base of van Fraassen's picture of science. It begins with the observation that at any given point we do not know whether some phenomenon will become observable in the fullness of time. Hence van Fraassen draws the observable/unobservable distinction for broadly epistemic reasons. He writes...

> The human organism is, from the point of view of physics, a certain kind of measuring apparatus. As such it has certain inherent limitations - which will be described in detail in the final physics and biology. It is these limitations to which the 'able' in 'observable' refers - our limitations, *qua* human beings.[59]

[57] [76]pg154

[58] [76]pg168

[59] [32]pg17.

So the idea here seems to be that there are objective possibilities regarding what we can and cannot come to see. The nature of what is observable is always relativised to us at any given point of evaluation. These objective possibilities support certain counterfactual claims. For instance, if the loch ness monster were to exist, we would be able to see it, but conversely if sub-sub-sub,...subatomic particles exist, we would not be able to see them. This kind of limitation underlies the cogency of the distinction and van Fraassen's modest empiricism.

There is a tension here since the motivation for the distinction was to defend empiricism (the view that we should repudiate belief in that which goes beyond our possible experience) and do away with extravagant and speculative metaphysics about modal facts of any stripe. Evidently the adoption of the objective distinction between observable and unobservable appears to commit van Fraassen to a view of counterfactual conditionals as conveying objective truth-claims. This is explicitly against his stated position. He claims that "the supposed objective modal distinctions drawn are but projected reifications of radically context dependent features of our language"[60] This tension suggests the ambiguity or incoherence of van Fraassen's position. James Ladyman goes to great pains to resolve any ambiguity and determines the position to be simply incoherent.[61]. The issues are mostly exegetical. For instance, he quotes van Fraassen making the following statement: "[T]o assert a theory is to assert that the actual, whatever it be shall fit (to a significant degree) the possibilities delimited by that theory. And I perceive no valid inference from this type of assertion to any form of realism with respect to possibilities or propensities".[62] Ladyman is rightly puzzled as to whether this means van Fraassen merely repudiates belief in modal commitments of a theory, or that he means to doubt that modal commitments are truth apt. The issue is resolved to be the former since Monton and van Fraassen partly concede this point in their reply to Ladyman.

We suggest that the lesson to draw here is that you could not be a constructive empiricist in van Fraassen's sense and abstain from a belief in objective modal facts about the nature of observability, since if you concede that there is nothing objective about observable/unobservable distinction, then what comes to count as an empirically adequate theory (based on observability criteria!) is effectively arbitrary depending on the local definition of observability. This is a perverse motivation for empiricism of any stripe - which should insist that our information comes from experience not *a priori* or from definition especially for the presumed empirical problem of our observational capacity. So either constructive empiricism is *ad hoc* or inconsistent. Van Fraassen and Monton reply to this charge by arguing that the apparently modal locutions in their theory are aids to inference stated in the language of the accepted theory, and that the observability notion does not encode an objective modal fact. Instead "observed" is a descriptive predicate nothing more. This response is largely uncompelling, since it tries to do away with the notion that science can provide objective counterfactual considerations about the limitations of observation. To assert that an object is observed is the same as saying it is solid, both properties are tested for under lab conditions. But no such experiments are thought to underwrite counterfactual considerations about when or why said object will be observed in the future. Odd as this sounds, it seems to indicate that van Fraassen concedes the point that constructive empiricism is, in this sense, *ad hoc* - based on a local definition (or test) for observability. They defend the move by stating it is a vulnerability of any non-foundational epistemology.[63] However, Ladyman's core point is that if the notion of observability cannot be used to underwrite counterfactual claims then we cannot use such counterfactuals to rule out the existence of unicorns and the loch ness monster, but this is precisely the function of such counterfactual conditionals in contemporary science. They underwrite the design of experimental tests, the placement of cameras, the nature of the equipment etc, etc.... The failure of these tests is only significant because of the presumed *observability* of

[60] Quoted in Ladyman's paper [57]pg847.

[61] cf. [57] and [58]

[62] See [57]pg847.

[63] cf [67]

such entities! We do not accept the existence of such monsters precisely because the experiments have failed to observe such creatures when counterfactual considerations would suggest that this is unlikely if in fact the loch ness monser were to exist. Hence, constructive empiricism seems, at best, to be a poor model of scientific practice. At worst it is a perverse empiricism with *ad hoc* restraints on the semantics of the observability notion.

Recap and Reflection

So for our current dialectic, we can conclude that constructive empiricism does not undermine our broadly realist theory with their insistence that the adequacy conditions on a scientific theory ought to be minimalist and constrained by empiricist motivations. We have distinguished between structural realism and constructive empiricism but despite the *prima facie* plausibility of the latter we have tried to show that neither our theory of explanation nor structural realism more generally is invalidated by the arguments of the constructive empiricist. As such we think that there is ample reason to accept the theory of explanation we have developed throughout this thesis. But we accept van Fraassen's central point - whether you believe the theory is another question.

6.7 Conclusion

In this chapter we have tried to (a) defend the role and importance of structural realism in the scientific enterprise, and (b) show how such structural information can be used to induce different epistemic and doxastic changes. By far we have spent most of our time developing (a) as a response to the challenge van Fraassen posed to structural and (more broadly) scientific realism. We deem this defence to be adequate. The trouble with developing (b) is that the topic is really deserving of a thesis on its own. At best our discussion has been suggestive about how to go about incorporating our structural information in the epistemological dynamics. We comfort ourselves with the idea that our suggestions are not entirely without merit. Indeed we take these considerations to be a promising start for future work.

Conclusions

The philosophical problem is an awareness of disorder in our concepts, and can be solved by ordering them. - Ludwig Wittgenstein in *The Big Typescript*

Introduction: Identifying Explanations

Consider the problem of John the arsenic eater. He accidentally consumes arsenic, and by any reasonable law relating to the effects of poison, we could predict his death. So on discovering that John has died, you might think to explain this event by appeal to his recent diet. However, if he also gets hit by a bus immediately after his recent meal, is your explanation a good one? Similarly, if John consumes birth control pills, can we really explain his failure to become pregnant by such a correlation? The challenge here is to find the appropriate conditions (both epistemic and ontic) which would allow ϕ to count as an explanatory answer to the why-ψ question. For problem cases, such as Scriven's famous paresis example, we do not know the full details of the causal structure which generates paresis. We only know, so Scriven reports, that it can develop in a small number of syphilitic patients. So does syphilis explain paresis, or should we postulate an unknown "disposition for paresis" and explain the paresis as conjunction of these factors? The arsenic eater-problem suggests that laws are too general and often too superficial to serve as an explanation in certain settings. The pregnancy-problem suggests that we must filter the appropriate kind of dependency relation from a slew of candidates, in particular we must develop a preference (or choice) rule of some kind to prioritise the appropriate candidate explanation. In both cases a more explicit focus on dependency relations is the only appropriate solution. More obviously, the paresis example shows a need for better understanding of the causal process underlying the events. So in any case there is a motivation for factoring directly for dependency structures in our explanatory reasoning. The broad moral here is that we should seek to identify explanatory claims with reference to the identified structures of dependency. Formalisms such as justification logic play the intermediary role of encoding the evidentiary relevance of the *explanans* to the *explanandum* just so long as the justificatory (or evidentiary) relations are defined with respect to observed (or implicitly acknowledged) dependencies.

Taxonomy of Alternatives

We briefly elaborate some historical alternatives to our model of explanations. But first consider this table. The idea here is that an explanation can be assessed the degree to which it addresses each of these aspects. Does it record a law like connection between the *explanans* and *explanadum*, does it effect our credences or make record of a structural dependency? Much of the history of theories of explanation in the philosophy of science has been spent on isolating one of these aspects in of explanatory claims.[64]

[64] The table is adapted from Salmon's discussion in [80]pg93

Aspect	Determinism	Indeterminism
Epistemic	Logical Necessity Nomic Subsumption Expectedness with Certainty	 High Inductive Probability Inductive Support High Nomic Expectability
Modal	Nomological Necessity/Possibility Lawful connection	 Only Statistical Modalities
Ontic	Intelligible structural Pattern Non-arbitrary Structural Dependence	 Intelligible structural pattern Pattern structured by probabilistic dependencies Probabilities need not be high

Hempel's D-N model

Most famously Carl Hempel invented the Deductive-Nomological account of explanation. The idea essentially states that an explanation is an argument from a universally quantified premise, in addition to certain contingent matters of fact. Pictorially:

$$\frac{\mathrm{L}_1 \ ... \ \mathrm{L}_n, \ \mathrm{C}_1 ... \mathrm{C}_n}{\mathrm{E}}$$

Each law L_i is universally quantified, so the *explanandum* is thought to be subsumed in the *explanans*. This presumes that the explanation relation is a kind of specialised entailment relation. Evidently this theory is too broad. On the one hand the theory is vulnerable to the Arsenic eater style objections, and on the other we have already seen that there are species of explanation which would not satisfy the properties predicted by Hempel's model of explanation due to the fact that are based on non-monotonic relations.

Hempel's I-S model

To deal with some of these difficulties Hempel adapts his theory to include a second paradigm of explanatory reasoning. Instead of deducing the *explanandum* from its *explanans* we need only determine that the *explanandum* is made highly probable by the *explanans*. In particular the probability of the explanatory inference is specified to be equal to the probability of the (necessarily high) conditional probability statement appearing in the *explanans*. However, the conclusion induced by the inference is always qualitative. So we inductively infer the unqualified truth of the conclusion on the basis of our probabilisitc dependency.

$$\frac{\sigma(\mathrm{F} \mid \mathrm{G}) = .95, \ \mathrm{G(a)}}{\mathrm{F(a)}} \ (.95)$$

This is an inductive argument, which insists that the inference to F(a), is unqualified given the high degree of probability of the conditional claim. Of course this setting is prone to validate explanations based on spurious associations, such as the supposed explanation of John's failed motherhood, due to his diet of pregnancy pills. Both of Hempel's models are inordinately focused on capturing the epistemic aspect of explanations, but they suffer in so far as the only models of reasoning which they assume underlie our epistemic notions are brute logical entailment and probabilistic correlation. We hope to have shown that this is too narrow a model of our epistemic and doxastic states.

Kitcher's Unification Model

This model of explanation focuses primarily on the role of explanation in the shaping of our epistemic states. Put more floridly, scientific explanations are compelling because "science advances our understanding of nature"[65] We are in agreement with Kitcher on this point but we have only suggested the details by which we link understanding and explanations. Systematising the impact of explanation on the growth of understanding requires, we think, a view of explanation as a species of information update. The focus here is primarily epistemic. Hence, our theory is at least compatible with the goals of Kitcher's project.

Salmon's Causal Model

Wesley Salmon once suggested that we "put the cause back into *because*".[66] The idea is that if we specify a causal requirement on certain explanations we could avoid certain instances of explanatory irrelevance and spurious association underlying the case of the arsenic eater and the pill popper. Our theory is perhaps closest in spirit, but different in detail, to this idea. Although some counterexamples have been discovered against this model, we hope you agree that this merely prompts issues about how to accurately model appropriate dependency relations e.g. the paresis example. The focus here is primarily ontic. But again, our theory is also compatible with the goals of Salmon's project.

Recap

In this thesis we have attempted to problematise the notion of rational explanation with underdetermination problems. In particular we motivated this problem by Goodman's grue paradox. We sought to address this picture by developing a model of rational inference capable of avoiding such underdetermination problems. We then showed the failure of the standard models of formal epistemology; in particular we showed that we could not recover a notion of rational belief on either the Bayesian or Hintikka models without amending them. We then proposed to augment our epistemological models with ideas from justification logic. The idea was that if we could specify beliefs as arising from a process of justification we could reject various underdetermination possibilities as unjustified, thereby avoiding grue-type problems.

Of course, then the natural question "whence justification?" arises. To address this concern we argued that we could specify justification logics in such a way that different species of justification could be seen as tracking different species of dependence relations. To make this clear we elaborated the logic of two such dependency relations (i.e. grounding, and causal dependence) so as to define justification logics based on these relations. We then applied these considerations to the resolution of Benacceraf's multiple reduction problem and Goodman's grue problem. Motivated by this result we then developed a theory of explanation as information update based on dependency structures. Note that the manner of our construction allows for multiple species of explanation based on the nature of the dependency claim each explanation encodes. This theory of explanation was observed to be viable only if the idea of incorporating actual dependency structures is sound. To make the case for this idea we defended structural realism against van Fraassen's minimalistic constructive empiricism.

Reflection: Gentle Polemics

Suppose you observe a man convinced of the unreality of things unobserved, walk into a transparent glass door. Would you expect him, like the fly, to continue unabashed instinctively bashing himself against the window in rank frustration with reality? If you're generous you expect him to

[65] [12]pg32.

[66] [66]pg787

concede the ontological claim that there exists an obstacle which has been unobserved, furthermore you expect him to now believe that there exist things that remain unobserved. Better, you expect him to open the door. This history of trail and error, which is the history of mankind, attests to the discovery of many such obstacles. Van Fraassen's apparent queasiness with the idea that such discoveries are even possible is indicative of a discomfort with the potential excess of a methodology which allows for overly speculative theorising about the nature of doorways. However, given the expectation that van Fraassen would open the glass door and walk through, we suggest that just the appropriate level of speculative reasoning was applied. If on the other hand van Fraassen sought to live by his own philosophical tenets, then we would expect fly-like behaviour and perform an exhaustive series of tests. We do not, so presumably no one can live by the law of constructive empiricism. It is a slightly different question as to whether scientists should enter into such speculative reasoning. But there is no real problem here. Ontological speculation is inevitable, science simply restrains inference by testing but hypothesis formation is based on speculation. Hence van Fraassen's worry about the ingress of harmful metaphysics is entirely unfounded. Similarly the voluntary focus of traditional formal epistemology is needlessly restrictive, thereby ensuring the inadequacy of those models.

We have only begun to examine the interplay between formal epistemology and formal ontology. In particular we owe an elaboration of how to systematically relate species of reasoning based on distinct dependency relations with different epistemological dynamics. There remain further kinds of explanatory update operations which we can and should relate to our reasoning about different species of dependence. Perhaps more importantly we should provide a ranking of which species of explanation supersede one another in a justification game. In general the social aspects of epistemology have been under examined in this thesis, but as it stands we deem our steps along this path to be promising. Our first test was successful. The theory of explanation we have developed has been shown (to be minimally) a viable alternative to existing theories, and in a large part more adaptable than those theories. This provides support for our criticism of the traditional models of formal epistemology and our dissatisfaction with them, while motivating the direction of our future work. However there are reasons to think that we are on the right path.

Conclusion

Unlike the traditional theories of explanation, our theory of explanation (even on the surface) incorporates both epistemic and ontic components. The implementation of such a theory requires the assumption of all three aspects of explanation adumbrated in the table above. This is opposed to traditional theories of explanation which seem to be crippled by either their narrow approach to capturing an aspect of explanation or their inattention to the fact that the dependency structures on which we reason are more varied than previously supposed. Our theory of explanation is flexible enough to accommodate any of the results of the traditional theories and avoid some of their deficits.

Secondly, many of the failures of the traditional models of formal epistemology are exhibited by noting their failure to articulate the problem of underdetermination. The extent to which we successfully address this issue, is due in a large part to the amendment to our epistemological doxastic models motivated by ideas underlying justification logic and the ability to capture procedures of rational inference required to address the underdetermination problem. We have tried to show that (a) models of knowledge and belief should be able to account for procedures of rational inference and (b) procedures of rational inference cannot be simply stipulated by fiat. Instead we chose to model distinct species of inference by ensuring that our models of reasoning are based on the appropriate models of dependence. Hence, the extent to which our theory of explanation is successful is based on a combination of formal epistemology and formal ontology. To paraphrase Kant - epistemology without ontology is empty, and ontology without epistemology is arbitrary.

Technical Appendix

The Realisation Theorem

Soundness and completeness results for KT4 can be found in most text books. So recall that the realisation theorem states that for any KT4 validity, we can recover a validity in JT4. In general this holds for any justification logic and the appropriate corresponding doxastic or epistemic logics. However, we are faced with a dilemma. The proof has been proven semantically with reference to a complex construction of a canonical model style of proof, and syntactically by a careful constructions of a proof tree. The syntactic proof is more intuitive, but to elaborate the core notions we have to introduce the proof theoretic system of sequent calculus.

Terminology

We define the notion of a tree structure.

Definition (Terminology for trees) Trees are partially ordered sets (X, $\leq$) with a lowest element and all sets of { y | y $\leq$ x } for x $\in$ X linearly ordered. The elements of X are called nodes, and branches are maximally linearly ordered subsets of X. The least node is called root of the tree. We call a tree *labelled* if each node is assigned a particular formula. The depth of a tree $\mid \mathcal{T} \mid$ is the maximum length of the branches in the tree. Labelled trees represent proof sequents with each node converging toward the root. That is to say we reason from the leaves of the tree to a derivation of the formula at the root.

We define the notion of a Formula Occurence in a proof sequent.

Definition (Terminology for Subformula Occurence) The definition is by induction:

- ϕ is a subformula of ϕ
- If $\psi \otimes \chi$ is a subformula of ϕ then so are ψ, χ, for $\otimes = \vee, \wedge, \rightarrow$
- If $\Box\psi$ is in ϕ, then so is ψ.

Definition (Positive, negative and strictly positive formulas)

- ϕ is a postive and strictly positive subformula of itself.
- If $\psi \wedge \chi$ or $\psi \vee \chi$ is a positive [negative, strictly positive] subformula of ϕ, then so are ψ, χ.
- If $\psi \rightarrow \chi$ is a positive [negative] subformula of ϕ then ψ is negative[positive] subformula of ϕ, and χ is positive [negative] subformula of ϕ. And if $\psi \rightarrow \chi$ is a strictly positive subformula of ϕ then so is χ.
- if $\Box\psi = \chi$ in ϕ, then the occurrence of $\Box$ in χ is positive. If $\chi[\neg\chi]$ occurs in either $(\tau \rightarrow \chi[\neg\chi])$, $(\tau \wedge \chi[\neg\chi])$, $(\chi[\neg\chi] \wedge \tau)$, $(\tau \vee \chi[\neg\chi])$, $(\chi[\neg\chi] \vee \tau)$, then the polarity of the $\Box$ in each is positive or negative respectively i.e. such constructions do not effect the polarity of $\chi[\neg\chi]$.

On the other hand the occurrence of $\Box$ in χ is altered if $\chi[\neg\chi]$ occurs in either $\neg(\chi[\neg\chi])$ or $\chi[\neg\chi] \to \tau$.

Sequent Calculus

The sequent calculus is a type of proof theory which is useful here because it allows us to reason about proofs at a meta-level. The core idea is that we incorporate the consequence relation $\vdash$ (represented here as $\Rightarrow$) in our proof system. A discussion of this motivation can be found in Troelstra and Schwichtenberg *Basic Proof Theory*[67] or Boolos *et al* in their *Computability and Logic*.[68] The basic structural and inferential rules for the propositional connectives are as expected. We wish to prove a result about a modal logic, so the extra rules for the modal consequence relation are as follows:

$$\frac{\mathrm{A}\ \Gamma \Rightarrow \Delta}{\Box \mathrm{A},\Gamma \Rightarrow \Delta}\,(\Box \Rightarrow) \qquad \frac{\Box\Gamma \Rightarrow \mathrm{A}}{\Box\Gamma \Rightarrow \Box \mathrm{A}}\,(\Rightarrow \Box) \qquad \frac{\mathrm{A}\Rightarrow \Diamond\Delta}{\Diamond \mathrm{A}\Rightarrow \Diamond\Delta}\,(\Diamond \Rightarrow)$$

$$\frac{\Gamma \Rightarrow \mathrm{A},\ \Delta}{\Gamma \Rightarrow \Diamond \mathrm{A},\ \Delta}\,(\Rightarrow \Diamond)$$

The idea is to specify that any valid sequence i.e. any constructable node on a tree in this system (S4 = KT4) is such that we can take each node of the tree and define a "realisation" function to recover a validity of justification logic i.e. a realisation of the premises on the left and a realisation of the consequences on the right. The $\Rightarrow \Box$ rule corresponds to the Necessitation rule in standard modal logic and the idea is that we have to recover the Internalisation rule. Hence the Lifting Lemma will have a crucial role in this proof. Since each node on a tree presents a valid sequent, each application of $\Rightarrow \Box$ presents a challenge for recovering an appropriate justification term. For each rank of the tree we must ensure that for each valid belief we can recover a justification of the formula believed. Consider the trees formed by this proof:

$$\text{(R}\to\text{)}\ \dfrac{\dfrac{\text{(L}\to\text{)}\ \dfrac{\text{(R}\to\text{)}\ \dfrac{\text{(RW)}\ \dfrac{\mathrm{A} \Rightarrow \mathrm{A}}{\mathrm{A} \Rightarrow \mathrm{A},\ \mathrm{B}}}{\Rightarrow \mathrm{A},\ \mathrm{A} \to \mathrm{B}} \qquad \dfrac{\dfrac{\dfrac{\mathrm{B} \Rightarrow \mathrm{B}}{\mathrm{B} \Rightarrow \mathrm{B},\ \neg \mathrm{A}}\,\text{(RW)}}{\mathrm{B},\ \neg \mathrm{B} \Rightarrow \neg \mathrm{A}}\,\text{(L}\neg\text{)}}{\mathrm{B},\ \Rightarrow \neg \mathrm{B}\to \neg \mathrm{A}}\,\text{(R}\to\text{)}}{\Rightarrow \mathrm{A}\to \mathrm{B},\ \mathrm{A}\to \mathrm{B} \Rightarrow \neg \mathrm{B} \to \neg \mathrm{A}}}{\Rightarrow \mathrm{A}\to \mathrm{B} \Rightarrow \neg \mathrm{B}\to \neg \mathrm{A}}\,\text{(C)}}{\Rightarrow (\mathrm{A}\to \mathrm{B}) \to (\neg \mathrm{B} \to \neg \mathrm{A})}$$

The proof is to be read from top to bottom, it shows that one direction of contraposition holds in this system. The other direction can be recovered similarly. The proof contains two branches. At each step we are very explicit about the nature of the rule being deployed. Since justification logic and doxastic logic coincide in terms the purely propositional inferences any tree not involving a modal claim constructed in S4 will be constructible in JT4 too. The interesting case involves proofs where the modal operator plays a role.

$$(\mathrm{R}\neg)\ \dfrac{\dfrac{(\Box \Rightarrow)\ \dfrac{\dfrac{A \Rightarrow A}{A, \neg A \Rightarrow}\,(\mathrm{L}\neg)}{A, \Box\neg A \Rightarrow}}{\Diamond A, \Box\neg A \Rightarrow}\,(\Diamond \Rightarrow)}{\Diamond A \Rightarrow \neg\Box\neg A}$$

The proof just shows that one side of the Duality holds. The other side is even simpler. This result shows that the the rules for $\Diamond$ can be dispensed with just so long as we take Duality to be definitional. The idea of the realisation theorem is that for each step on the S4 proof tree we can define a CS function such that each modal claim finds a justification, and the consequences of those modal claims are similarly justified.[69]. To do this properly we should indicate how justification

[67] cf. [91]

[68] cf. [15]

[69] Fuller details of sequent calculi can be found in [91] and [15]

logic is also to be specified in a sequent calculi. The propositional rules are the same as in S4. Differences arise only for inferences containing justification terms:

$$\frac{\mathrm{A}, \Gamma \Rightarrow \Delta}{\mathrm{j{:}A}, \Gamma \Rightarrow \Delta}\,(:\Rightarrow) \quad \frac{\Gamma \Rightarrow \Delta, \mathrm{j{:}\ A}}{\Gamma \Rightarrow \Delta, \mathrm{j!{:}j{:}A}}\,(\Rightarrow!) \quad \frac{\Gamma \Rightarrow \Delta, \mathrm{j{:}A}}{\Gamma \Rightarrow \Delta, \mathrm{(j{+}e){:}A}}\,(\Rightarrow +)$$

$$\frac{\Gamma \Rightarrow \Delta, \mathrm{j{:}A{\to}B} \qquad \Gamma \Rightarrow \Delta\ \mathrm{e{:}A}}{\Gamma \Rightarrow \Delta, \mathrm{(j \cdot e){:}B}}\,(\Rightarrow \cdot) \quad \frac{\Gamma \Rightarrow \mathrm{Ax}\ ,\Delta}{\Gamma \Rightarrow \mathrm{j{:}Ax}\ \Delta}\,(\Rightarrow\mathrm{j})$$

It is trivial to prove the **factivity** axiom by succesive applications of (:⇒) and (R→). More clearly:

$$\dfrac{(:\Rightarrow)\ \dfrac{A\Rightarrow A}{j : A\Rightarrow A}}{\Rightarrow j : A \to A}\,(\mathrm{R}\to)$$

We say that the justification logic $\mathrm{JT4}_0$ is characterised by the propositional structural rules of S4 and the rules: (:⇒), (⇒!), (⇒ +) and (⇒ ·). This is just to ensure that all the axioms are validated. Adding (⇒j) is akin to encoding the **axiom necessitation** rule, and hence JT4 is specified in the sequent calculus by adding these five rules to the rules of the classical propositional calculus. As we've already shown the assumption of **axiom necessitation** allows us to prove the Lifting Lemma. Artemov proves a soundness and completeness result for the sequent calculus JT4 by the usual method.[70]

The Theorem

It now remains to sketch the construction involved in the proof.

Definition (A Realisation modulo CS) By a realisation of ϕ in S4 we mean an assignment of justification terms to to all occurrences of $\Box \in \phi$. Denoted ϕ^r. A realisation is *normal* if all negative occurrences of □ are assigned a justification variable, and the CS-function is injective i.e. each justification constant is assigned to only one formula.

Theorem (The Realisation Theorem) If S4 $\vdash \phi$, then JT4 $\vdash \phi^r$ for some normal realisation.

Proof: If S4 $\vdash \phi$ then there exists a derivation tree $\mathcal{T}$ of a sequent $\Rightarrow \phi$. It suffices now to construct a normal realisation r with an injective constant specification CS such that $\mathrm{JT4}_{CS} \vdash \bigwedge \Gamma^r \to \bigvee \Delta^r$ for any sequent $\Gamma \Rightarrow \Delta$ in $\mathcal{T}$.

The rules of the S4 sequent calculus are such that the respect the polarities of as expected. In addition to the definition above we can see that all occurences of □ introduced by (⇒ □) are positive, while negative occurences of □ are introduced by either (□ ⇒) or weakening.

We call two occurrences of □ *related* if they occur in a premise and consequent determined by a particular inference or structural rule. We extend this relationship by transitivity along the branches of our trees. If branches diverge then we can split our tree into disjoint, so to speak, families of ancestors and descendants. We call a family *essential* if at least one instance of the (⇒ □) rule occurs. Retaining the familial metaphor we can think of this relation as indicating a closeness amongst the family not found in others. Consider the proof the following claim: ⇒ (□A ∨ □B) → □(□A ∨ □B).

[70] See [4] for the proof itself and [15] for a discussion of the method.

$$\begin{array}{rcl}
 & \mathrm{A} \Rightarrow \mathrm{A} & \\ \hline
(\Box \Rightarrow) & \Box \mathrm{A} \Rightarrow \mathrm{A} & \\ \hline
 & \Box \mathrm{A} \Rightarrow \Box \mathrm{A} & (\Rightarrow \Box) \\ \hline
(\mathrm{RW}) & \Box \mathrm{A} \Rightarrow \Box \mathrm{A}, \Box \mathrm{B} & \\ \hline
 & \Box \mathrm{A} \Rightarrow \Box \mathrm{A} \vee \Box \mathrm{B} & (\mathrm{R}\vee) \\ \hline
(\Rightarrow \Box) & \Box \mathrm{A} \Rightarrow \Box(\Box \mathrm{A} \vee \Box \mathrm{B}) & \\ \hline
(\mathrm{RW}) & \Box \mathrm{A} \Rightarrow \Box(\Box \mathrm{A} \vee \Box \mathrm{B}), \Box \mathrm{B} & \\ \hline
 & \Box \mathrm{A} \vee \Box \mathrm{B} \Rightarrow \Box(\Box \mathrm{A} \vee \Box \mathrm{B}) & (\mathrm{L}\vee) \\ \hline
(\mathrm{R}\rightarrow) & \Rightarrow \Box \mathrm{A} \vee \Box \mathrm{B} \rightarrow \Box(\Box \mathrm{A} \vee \Box \mathrm{B}) &
\end{array}$$

Note that there are no branches and the $(\Rightarrow \Box)$ rule is applied at the second step. By transitivity of the *essential* connection we say that this proof displays the ancestry of an *essential* family. Secondly there are only two usages of the $(\Rightarrow \Box)$ rule.

The Construction

First we distinguish between justification variables, provisional justifications, and actual justification terms. There are three steps to this procedure.

Step1 For any negative or non-essential positive families i.e. those family trees with negative occurrences of $\Box$, or occurrences of $\Box$ which are not introduced by the $\Rightarrow \Box$ rule. For any such families we replace each occurrence of $\Box$, with a proof variable e.g. $\Box\phi$ becomes x:ϕ.

Step2 Now for any *essential* family f we enumerate each occurrence of $\Rightarrow \Box$ in the proof tree. We allow f_n to be the number of such occurrences in the family. Replace all occurrences $\Box_i$ in our proof tree $\mathcal{T}$ with the sum of provisional justification terms e_i i.e. replace each $\Box_i$ with $(\mathrm{v}_1+\mathrm{v}_2... +\mathrm{v}_{f_n})$ up until a maximal point of $\Box_{f_n}$. This procedure results in a proof tree $\mathcal{T}$' of the right shape for JT4. It remains to show that the proof tree is in fact valid in JT4.

Step3 For this step we need to replace the provisional justification terms in such a way that the validity of each step in the tree is preserved. The proof is by induction beginning at the highest leaves on the tree - the idea is to find a CS function so that we can find novel justification constants to justify each proposition by $(\Rightarrow \Box)$. To begin with the CS function is empty.

Base Case: A $\Rightarrow$ A is valid in both JT4 and S4 for any A. Hence, the highest leaf in an S4 tree can be recovered in any JT4 tree.

Induction Hypothesis For all nodes lesser than k on an S4 tree $\mathcal{T}$, there is a corresponding sequences of nodes in a in a JT4 tree $\mathcal{T}$' where all formulas of the form $\Box\phi$ in the S4 tree are realised in line with our stepwise construction, and there is a valid derivation for each inference step up to k-1.

Propositional Rules For all propositional rules we do not change our CS function. The application of each such rule can be repeated in JT4 since both are classical languages sharing the same sequent calculus for all non-modal cases.

$(\Rightarrow \Box)$ A node corresponding the usage of this rule will have the following shape.

$$\dfrac{\Box\Gamma \Rightarrow \mathrm{A}}{\Box\Gamma \Rightarrow \Box \mathrm{A}}\ (\Rightarrow \Box)$$

Allowing that Γ is a multiset and that $\Box \Rightarrow$ is applied. We have by our **Step1** we have that:

$$\dfrac{\mathrm{y{:}B}_1,\ \mathrm{y{:}B}_2....\ \mathrm{y{:}B}_n \Rightarrow \mathrm{A}}{\mathrm{y{:}B}_1,\ \mathrm{y{:}B}_2....\ \mathrm{y{:}B}_n \Rightarrow}$$

The challenge is to show that there is a justification term for A. The idea is to take the sum of the all justification terms for each instance of □ introduced by ⇒ □, by the monotonicity of such terms this will serve as a justification for A. Our intent is to introduce justification constants for each application of ⇒ □. By construction all $y_1....y_n$ are justification variables and each of $e_1....e_{f_n}$ is a provisional justification.. Hence by induction hypothesis...

$$\dfrac{y{:}B_1,\ y{:}B_2....\ y{:}B_n \Rightarrow A}{y{:}B_1,\ y{:}B_2....\ y{:}B_n \Rightarrow (e_1+\ .....\ +e_{f_n}) : A}$$

...with e_i corresponding to the i-th application of ⇒ □. Let e_m with $(1 \leq m \leq f_n)$ be a provisional justification for the particular claim A.

Now we apply the Lifting Lemma[71] to (a) create a new justification term and (b) extend our CS function to ensure that such a term will remain relevant to A. This is the crucial step. By assumption $y{:}B_1$, $y{:}B_2....$ $y{:}B_n \vdash A$, and so by the lifting lemma there is a constant justification term, denoted $t(y_1\ ...\ y_n)$ such that $t(y_1\ ...\ y_n){:}A$. By the monotonicity of justification we know that:

$$\begin{gathered} t{:}A \rightarrow (e_1+.....+e_{m-1} + t + e_{m+1}+...+ e_{n_f}){:}A).\text{ and hence} \\ \vdash y{:}B_1,\ y{:}B_2....\ y{:}B_n \Rightarrow (e_1+.....+e_{m-1} + t +e_{\ m+1}+...+ e_{n_f}){:}A \end{gathered}$$

Now that we have found an appropriate justification term, we shall replace all instances of e_m in the proof tree $\mathcal{T}$' and CS. This completes the induction step. Furthermore, because we have replaced all instances of e_m with t our CS function remains injective.

Repeating this procedure for each provisional justification term e_i will yield a proof tree $\mathcal{T}$" such that for every instance of ⇒ □ we have replaced □ with a justification constant, and the step is valid in JT4. ⊣.

An Example

So for instance, consider for the proof above of □A ∨ □B ⇒ □(□A ∨ □B). We display one possible realisation of this proof. There are other proofs which have alternative realisations.[72] It's perhaps simpler to see that the first application of ⇒ □ in (□A ⇒ □A), gets realised as $(x{:}\ A) \rightarrow (j \cdot x{:}\ A)$ where j is the constant justifying the axiom $(A \rightarrow A)$. Assuming **factivity**, A follows by modus ponens, so by the lifting lemma we have found a particular justification constant $(j \cdot x)$. Similarly, the other case of ⇒ □, we suggest that one realisation of the is derived from the combination of justification $r{:}(\phi \rightarrow \phi \vee \psi)$, for the classical axiom with the justification $!(j \cdot x)$, allowing us to derive, $(x : A)$ and hence also the desired disjunction.

$$(R\!\rightarrow)\ \dfrac{\dfrac{(RW)\ \dfrac{\dfrac{(\Rightarrow\Box)\ \dfrac{\dfrac{(RW)\ \dfrac{\dfrac{\dfrac{(\Box\Rightarrow)\ \dfrac{A \Rightarrow A}{(x){:}\ A \Rightarrow A}}{(x){:}\ A \Rightarrow (e_1+e_2){:}\ A}\ (\Rightarrow\Box)}{(x){:}\ A \Rightarrow (j\cdot x + e_2){:}\ A}\ (\text{Lifting})}{(x) : A \Rightarrow (j\cdot x + e_2){:}\ A,\ (y) :B}}{(x) : A \Rightarrow ((j\cdot x + e_2){:}\ A \vee (y) :B)}\ (R\vee)}{(x) : A \Rightarrow (j\cdot x + e_2){:}\ ((j\cdot x + e_2){:}\ A \vee (y) :B)}}{(x) : A \Rightarrow (j\cdot x + r\cdot !(j{\cdot}x)){:}\ ((j\cdot x + r{\cdot}!(j{\cdot}x)) : A \vee (y) :B)}\ (\text{Lifting})}{(x) : A \Rightarrow (j\cdot x + r\cdot !(j{\cdot}x)){:}\ ((j\cdot x + r{\cdot}!(j\cdot x) : A) \vee (y) :B),\ (y) : B}}{((x) : A \vee (y) : B) \Rightarrow (j\cdot x + r\cdot !(j{\cdot}x)){:}\ ((j\cdot x + r{\cdot}!(j{\cdot}x) : A \vee (y) :B)}\ (L\vee)}{\Rightarrow ((x) : A \vee (y) : B) \rightarrow (j\cdot x + r\cdot !(j{\cdot}x)){:}\ ((j\cdot x + r{\cdot}!(j\cdot x) : A \vee (y) :B)}$$

[71] We proved this earlier in our third chapter

[72] See the discussion in Artemov's paper [4]

This concludes one normal realisation of the above proof. Other proofs of the same theorem will receive a different realisation. The final justification term can be neatened if we trim the redundant conjuncts joined by the + operation. This is possible by the nature of our construction. Taking the realisation theorem together with the preservation theorem for KT4 proved in chapter three, we can conclude with this corollary.

Corollary KT4 $\vdash \phi \Leftrightarrow$ JT4 $\vdash \phi^r$ for some normal realisation.

This result shows formally, the deep connection between belief and the process of discovery which underlies such beliefs, which we have argued throughout this thesis.

Bibliography

[1] Jonathan Eric Adler and Lance J. Rips. *Reasoning: Studies of Human Inference and its Foundations.* Cambridge University Press, 2008.

[2] Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2):510–530, 1985.

[3] H. Arlo-Costa and J. Helzner. Ambiguity aversion: the explanatory power of indeterminate probabilities. *Synthese*, 2010.

[4] Sergei Artemov. Explicit provability and constructive semantics. *The Bulletin for Symbolic Logic*, 2001.

[5] Sergei Artemov. The logic of justification. *Review of Symbolic Logic*, 1(4):477–513, 2008.

[6] Sergei Artemov and Rosalie Iemhoff. The basic intuitionistic logic of proofs. *Journal of Symbolic Logic*, 72(2):439–451, 2007.

[7] A. Baltag, B.Renne, and S.Smets. The logic of justified belief change, soft evidence and defeasible knowledge. *Lecture Notes in Computer Science*, 7456:pp 168–190, 2012.

[8] A. Baltag and S. Smets. The logic of conditional doxastic actions. In *Texts in Logic and Games.* Amsterdam University Press, 2008.

[9] A. Baltag and S. Smets. A qualitative theory of dynamic interactive belief revision. In *Logic and the Foundations of Game and Decision Theory.* Amsterdam University Press, 2008.

[10] A. Baltag and S. Smets. Keep changing your beliefs, aiming for the truth. *Erkenntnis*, 75(2):255–270, 2011.

[11] Prasanta S. Bandyopadhyay and Malcolm Forster. *Philosophy of Statistics, Handbook of the Philosophy of Science, Volume 7.* Elsevier, forthcoming.

[12] Robert W. Batterman. *The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence.* Oxford University Press, 2002.

[13] Paul Benacerraf. What numbers could not be. *The Philosophical Review*, 74(1):47–73, 1965.

[14] Paul Benacerraf. Mathematical truth. *The Journal of Philosophy*, 70(19):661–679, 1973.

[15] George Boolos, John Burgess, Richard P., and C. Jeffrey. *Computability and Logic.* Cambridge University Press, 2007.

[16] C. Boutilier and V.Becher. Abduction as belief revision. *Artificial Intelligence*, 1995.

[17] B.Renne. *Dynamic Epistemic Logic with Justification.* PhD thesis, CUNY Faculty of Computer Science., 2008.

[18] Ross .P. Cameron. Turtles all the way down: Regress priority, and fundamentality. *The Philosophical Quartlerly*, 2008.

[19] A. Chalmers. Drawing philosophical lessons from perrin's experiments on brownian motion: A response to van fraassen. *British Journal for the Philosophy of Science*, 62(4):711–732, 2011.

[20] David Chalmers. *Constructing the World.* Oxford University Press, 2012.

[21] David John Chalmers, David Manley, and Ryan Wasserman. *Metametaphysics: New Essays on the Foundations of Ontology.* Oxford University Press, 2009.

[22] Silvio Seno Chibeni. Explanations in microphysics: A response to van fraassen's argument. *Principia*, 12(1):49–72, 2008.

[23] Michael Clark and David Liggins. Recent work on grounding. *Analysis*, 0, 2012.

[24] Boudewijn de Bruin. *Explaining Games: The Epistemic Programme in Game Theory.* Springer, 2010.

[25] Igor Douven. A principled solution to fitch's paradox. *Erkenntnis*, 62(1):47–69, 2005.

[26] Igor Douven. Abduction. In Edward N Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Stanford University, 2011.

[27] John Earman. *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory.* The MIT Press, 1992.

[28] Kit Fine. The logic of pure ground. *The Review of Symbolic Logic*, 2011.

[29] M.C. Fitting. The logic of proofs, semantically. *Annals of Pure and Applied Logic*, 2004.

[30] M.C. Fitting. Explicit logics of knowledge and conservativity. In *Proceedings, Tenth International Symposium on Artificial Intelligence and Mathematics*, 2008.

[31] Richard Foley. Beliefs, degrees of belief, and the lockean thesis. In *Degrees of Belief*, volume 342, pages 37–47. Springer, 2009.

[32] Bas C. Van Fraassen. *The Scientific Image.* Oxford University Press, 1980.

[33] Bas C. Van Fraassen. *Laws and Symmetry.* Oxford University Press, 1989.

[34] R. Frigg and I. Votsis. Everything you always wanted to know about structural realism but were afraid to ask. *European Journal of the Philosophy of Science*, 1:227–276, 2011.

[35] P. Gardenfors. *Knowledge in Flux: Modelling Dynamics of Epistemic States.* MIT Press, 1988.

[36] Dan Geiger. *Graphoids: A Qualitative Framework for Probabilistic Inference.* PhD thesis, UCLA, 1990.

[37] D Geiger T, Verma. Identifying independence in bayesian networks. *NETWORKS*, 1990.

[38] Jelle Gerbrandy. *Bisimulations on Planet Kripke.* PhD thesis, University of Amsterdam, ILLC, 1998.

[39] Clark Glymour. Relevant evidence. *Journal of Philosophy*, 72(14):403–426, 1975.

[40] Clark Glymour, P Sprites, and R. Scheines. *Causation, Prediction and Search.* Springer Verlag, 1993.

[41] Nelson Goodman. *Ways of Worldmaking.* Harvester Press, 1978.

[42] Nelson Goodman. *Fact, Fiction, and Forecast.* Harvard University Press, 1983.

[43] Bob Hale and Aviv Hoffmann. *Modality: Metaphysics, Logic, and Epistemology.* Oxford University Press, 2010.

[44] Joseph Y. Halpern. Axiomatising causal reasoning. *Journal of A.I Research*, pages 317–337, 2000.

[45] Joseph. Y. Halpern. *Reasoning about Uncertainty.* The MIT Press, 2003.

[46] Joseph Y. Halpern and Christopher Hitchcock. Actual causation and the art of modelling. In R Dechter, H Geffner, and Joseph Y. Halpern, editors, *Heuristics, Probability and Causality.* College Publications, 2010.

[47] Joseph Y. Halpern and Judea Pearl. Causes and explanations: A structural-model approach. part i: Causes. *British Journal for the Philosophy of Science*, 56(4):843–887, 2005.

[48] Joseph.Y. Halpern and Christopher Hitchcock. Graded causation and defaults. Unpublished Manuscript.

[49] J. Hintikka. *Knowledge and Belief: An Introduction to the Logic of the Two Notions.* Cornell University Press, 1964.

[50] Thomas Jech. *Set Theory.* Springer, 3rd edition, 2006.

[51] C. S. Jenkins. Is metaphysical dependence irreflexive? *The Monist*, 94(2):267–276, 2011.

[52] Kevin Kelly. A close shave with realism: How ockham's razor helps us find the truth.

[53] Kevin Kelly. *The Logic of Reliable Inquiry.* Oxford University Press, USA, 1996.

[54] Trogdon Kelly. Grounding: Necessary or contingent? *Pacific Philosophical Quarterly*, forthcoming.

[55] Philip Kitcher and Wesley Salmon. Van fraassen on explanation. *Journal of Philosophy*, 84(6):315–330, 1987.

[56] Saul A. Kripke. *Philosophical Troubles. Collected Papers Vol I.* Oxford University Press, 2011.

[57] James Ladyman. What's really wrong with constructive empiricism? van fraassen and the metaphysics of modality. *British Journal for the Philosophy of Science*, 51(4):837–856, 2000.

[58] James Ladyman. Constructive empiricism and modal metaphysics: A reply to monton and van fraassen. *British Journal for the Philosophy of Science*, 55(4):755–765, 2004.

[59] L.A.Paul. What mary can't expect when she's expecting. *Res Philosophica*, Forthcoming.

[60] Keith Lehrer. *Theory of Knowledge.* Westview Press, 2000.

[61] Keith Lehrer and Thomas Paxson Jr. Knowledge: Undefeated justified true belief. *Journal of Philosophy*, 66(8):225–237, 1969.

[62] David K. Lewis. *Papers in Metaphysics and Epistemology.* Cambridge, Uk ;Cambridge University Press, 1999.

[63] David K. Lewis. *Counterfactuals.* Blackwell Publishers, 2001.

[64] Peter Lipton. *Inference to the Best Explanation.* Routledge/Taylor and Francis Group, 2004.

[65] Studeny M. Multiinformation and the problem of the characterisation of the independence relation. *Problems of Control and Information Theory*, 3, 1989.

[66] J.A. Cover. Martin Curd, editor. *The Philosophy of Science: The Central Issues.* Norton, 1998.

[67] Bradley Monton and Bas C. van Fraassen. Constructive empiricism and modal nominalism. *British Journal for the Philosophy of Science*, 54(3):405–422, 2003.

[68] Ilkka Niiniluoto. Revising beliefs towards the truth. *Erkenntnis*, 75(2):165–181, 2011.

[69] O.Board. Dynamic interactive epistemology. *Games and Economic Behaviour*, 49(49-80), 2004.

[70] Martin Osborne and Ariel Rubinstein. *A Course in Game Theory.* The MIT Press, first edition, 1994.

[71] Andres Paez. Artificial explanations:the epistemological interpretation of explanation in ai. *Synthese*, 170(1):115–131, 2009.

[72] Maarten de Rijke Patrick Blackburn, Yde Venema. *Modal Logic.* Cambridge University Press, 2002.

[73] Judea Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, 2000.

[74] W. V. Quine. *Quintessence: Basic Readings From the Philosophy of W.V. Quine.* Belknap Press of Harvard University Press, 2004.

[75] Michael J. Raven. Is ground a strict partial order? *American Philosophical Quarterly*, 50(2):191–199, 2013.

[76] Gideon Rosen. What is constructive empiricism? *Philosophical Studies*, 74(2):143–178, 1994.

[77] Hans Rott. Stability, strength and sensitivity:converting belief into knowledge. *Erkenntnis*, 61(2-3):469–493, 2004.

[78] Darrell P. Rowbottom. Evolutionary epistemology and the aim of science. *Australasian Journal of Philosophy*, 88(2):209–225, 2010.

[79] Joe Salerno. *New Essays on the Knowability Paradox.* Oxford University Press, 2009.

[80] Wesley C. Salmon. *Causality and Explanation.* Oxford University Press, 1998.

[81] Jonathan Schaffer. Is there a fundamental level? *Noûs*, 2003.

[82] Wilfrid S. Sellars. Empiricism and the philosophy of mind. *Minnesota Studies in the Philosophy of Science*, 1:253–329, 1956.

[83] Theodore Sider. *Writing the Book of the World.* Oxford University Press, 2011.

[84] F. Soler-Toscano and F.R. Velazquez-Quesada. A dynamic epistemic approach to abductive reasoning. In *Dialogues and the Games of Logic: A Philosophical Perspective.* College Publications, London, 2012.

[85] Wolfgang Spohn. Stochastic independence, causal independence, and shieldability. *Journal of Philosophical Logic*, 9(1):73–99, 1980.

[86] Robert Stalnaker. *Inquiry.* Cambridge University Press, 1984.

[87] Robert Stalnaker. The problem of logical omniscience, i. *Synthese*, 89(3):425–440, 1991.

[88] Robert Stalnaker. Varieties of supervenience. *Philosophical Perspectives*, 10:221–42, 1996.

[89] Eric Steinhart. Why numbers are sets. *Synthese*, 133(3):343–361, 2002.

[90] Keith Stenning, Michiel van-Lambalgen. *Human Reasoning and Cognitive Science*. MIT Press, 2008.

[91] A.S. Troelstra and H. Schwichtenberg. *Basic Proof Theory*. Cambridge University Press, 2000.

[92] Johan van Benthem. *Logical Dynamics of Information and Interaction*. Cambridge University Press, 2011.

[93] Hans van Ditmarsch, Wiebe van der Hoek, and Barteld Kooi. *Dynamic Epistemic Logic*. Springer, 1st edition, 2007.

[94] Frank Veltman. Defaults in update semantics. *Journal of Philosophical Logic*, 25(3):221–261, 1996.

[95] Tom S. Verma and Judea Pearl. Causal networks: Semantics and expressiveness. 04 2013.

[96] Crispin Wright. *Truth and Objectivity*. Harvard University Press, 1992.