The London School of Economics and Political Sciences

Conceptualizing Uncertainty

The IPCC, Model Robustness and the Weight of Evidence

Margherita Harris

A thesis submitted to the Department of Philosophy, Logic and Scientific Method of the London School of Economics for the degree of Doctor of Philosophy, London, October 2021

Declaration

The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. In accordance with the Regulations, I have deposited an electronic copy of it in LSE Theses Online held by the British Library of Political and Economic Science and have granted permission for my thesis to be made available for public reference. Otherwise, this thesis may not be reproduced without my prior written consent.

I confirm that Section 4.2 is to be published in *Synthese* (Harris, 2021)

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of 92,254 words.

Margherita Harris

Abstract

The aim of this thesis is to improve our understanding of how to assess and communicate uncertainty in areas of research deeply afflicted by it, the assessment and communication of which are made more fraught still by the studies' immediate policy implications. The IPCC is my case study throughout the thesis, which consists of three parts. In Part 1, I offer a thorough diagnosis of conceptual problems faced by the IPCC uncertainty framework. The main problem I discuss is the persistent ambiguity surrounding the concepts of 'confidence' and 'likelihood'; I argue that the lack of a conceptually valid interpretation of these concepts compatible with the IPCC uncertainty guide's recommendations has worrying implications for both the IPCC authors' treatment of uncertainties and the interpretability of the information provided in the AR5. Finally, I show that an understanding of the reasons behind the IPCC's decision to include two uncertainty scales can offer insights into the nature of this problem. In Part 2, I review what philosophers have said about model-based robustness analysis. I assess several arguments that have been offered for its epistemic import and relate this discussion to the context of climate model ensembles. I also discuss various measures of independence in the climate literature, and assess the extent to which these measures can help evaluate the epistemic import of model robustness. In Part 3, I explore the notion of the 'weight of evidence' typically associated with Keynes. I argue that the Bayesian (or anyone who believes the role of probability in inductive inference is to quantify the degree of belief to assign to a hypothesis given the evidence) is bound to struggle with this notion, and draw some lessons from this fact. Finally, I critically assess some recent proposals for a new IPCC uncertainty framework that significantly depart from the current one.

To my mum, who would have given me the biggest hug of all.

Acknowledgements

Spending the last eighteen months stuck at home during a worldwide pandemic is certainly not how I imagined completing my thesis. However, it is especially in these times one realizes how lucky one is to have been granted the (if only temporary) stability that PhD funding affords one to keep thinking about what one wants to think about, despite the world outside looking so grim. So my thanks go, first of all, to LSE for giving me a chance to think!

Very many thanks to my supervisor Roman Frigg, for the invaluable support, trust and encouragement he has offered me whilst writing this thesis. And to my secondary supervisor Liam Kofi Bright, for all his insightful comments and stimulating conversations.

Thanks to all the regular participants of the working models reading group, for all the lively discussions and their invaluable company during this pandemic: Michal Hladky, Anatolii Kozlov, Insa Lawler, Dan Li, James Nguyen, Phillip Verreault Julien, Sam Rijken, Joe Roussos, Lorenzo Sartori, Rawad El Skafwhich and Mike Stuart. Special thanks to Joe, for all the helpful chats throughout this PhD; to James, for the endless and endlessly fun philosophical disputes; and to Michal, who kept this reading group going no matter what.

Thanks to Deborah Mayo for opening my eyes to how far "the statistics wars" are from being over, and most of all for being a source of inspiration.

Thanks to my LSE PhD peers and all the LSE academics who have made the last four years a very precious experience.

Thanks to my dad, for being calm in the storm and for teaching me a good few lessons over the years. And to Matt, for his friendship, love and immense support all these years.

Finally, thanks to all the LSE builders, who have worked tirelessly from the beginning to the very end of the writing of this thesis. Keeping up with their constant incremental progress on the Marshall Building from the high floors of LSE's PhD Academy has been by far my healthiest procrastination routine whilst writing this thesis.

Contents

Introduction **1**

I An assessment of the IPCC conceptualization of uncertainty **7**

1 Some conceptual problems in the IPCC **8**

1.1 Introduction 8

1.2 The current IPCC uncertainty framework (for the AR5 and the AR6) 10

1.3 The perplexing bifurcation between evidence and agreement in the characterization of confidence 15

1.4 What types of uncertainty do confidence and likelihood represent? 20

1.4.1 No interpretation fits the bill 25

1.5 Troubling implications 28

1.5.1 The lack of transparency behind the interaction between confidence and likelihood levels 28

1.5.2 Value judgments and non-interpretable findings 31

1.6 Taking Stock 32

2 A genealogy of 'confidence' and 'likelihood' and what we can learn from it **35**

2.1 Introduction 35

2.2 The first IPCC uncertainty framework (for the AR3) 37

2.3 The second IPCC uncertainty guide (for the AR4) 42

2.4 On the reason(s) for the emergence of two uncertainty scales . . . 47

2.5 Likelihood revisited: objective probabilities, subjective probabilities or neither? 51

2.5.1 A Review of Some Statistical Concepts 51

2.5.2 A closer look at likelihood: “multi-model ensemble methods" and a questionable desire for “objectivity". 60
2.6 Taking stock 73

II Model-based robustness analysis 77

3 Robustness analysis as tool for discovering robust theorems 78
3.1 Introduction 78
3.2 Robustness reasoning “in action” 87
3.3 Robust theorems, low-level confirmation and ceteris paribus clauses 96
3.3.1 Weisberg’s general characterization of robustness analysis 97
3.3.2 On the empirical content of robust theorems 102

4 Robustness analysis as a tool for confirming robust theorems: an assessment of some popular arguments 113
4.1 Introduction 113
4.2 The epistemic value of independent lies: false analogies and equivocations. 114
4.2.1 An argument from coincidence? 118
4.2.2 What is Kuorikoski et al.’s argument? 125
4.2.3 A prima-facie more plausible argument (and yet...) 130
4.3 A critical assessment of Schupbach’s explanatory account of model-based robustness analysis 138
4.3.1 Schupbach’s explanatory account of RA diversity 140
4.3.2 Empirically driven RAs 143
4.3.3 Does Schupbach’s account of ERA diversity apply to model-based RA? 146

5 The epistemic import of model agreement in climate science: what philosophers and scientists have to say about it 157
5.1 Introduction 157
5.2 Lloyd and Parker on the epistemic import of model robustness in climate science 160

5.3 Winsberg on ERA diversity and Climate model ensembles 170
5.4 Independence revisited: what have climate scientists said about the epistemic import of model agreement? 180
5.4.1 Measures of independence: A-posteriori and A-priori approaches 184
5.4.2 Why such a strong focus on independence? 189

III The weight of evidence and some proposals for a new IPCC uncertainty framework 200

6 On the weight of evidence: what is it, can we measure it, and why should care about it? 201
6.1 Introduction 201
6.2 Keynes on the weight of arguments: two unmeasurable concepts 204
6.3 The Bayesian on the weight of evidence 218
6.3.1 The weight of evidence and severity: two (very different) sides of the same coin? 234

7 An assessment of some proposals for a new IPCC uncertainty framework 238
7.1 Introduction 238
7.2 Winsberg's proposal: Can scientists measure the resilience of their credences, and should they? 240
7.3 Mach et al.'s proposal: So long confidence? 262
7.4 Some thoughts towards an adequate uncertainty framework: Avoiding the same old mistakes 266
7.5 Bradley et al.'s proposal: What decision makers want . . . and how to give it to them without being peer pressured 275

Concluding Remarks 285

Bibliography 291

Introduction

The aim of this thesis is to improve our understanding of how to assess and communicate uncertainty in areas of research deeply afflicted by it, and where the assessment and communication of that uncertainty are made fraughter still by the studies' immediate policy implications. The IPCC is my case study throughout the thesis, which consists of three parts.

In the first part of my thesis, I offer a thorough diagnosis of some conceptual problems faced by the IPCC uncertainty framework. This I do because I believe that any successful attempt to revise and improve this framework will have to start from a clear understanding of the conceptual problems it currently faces, their implications for the IPCC authors' treatment of uncertainties, and the quality of the information provided in the IPCC uncertainty report. Accordingly, Part 1 sets out to contribute to this first step. It consists of the following two chapters.

In Chapter 1, I discuss two important conceptual problems in the current IPCC uncertainty framework: the puzzling bifurcation between evidence and agreement in the characterization of 'confidence'; and the lack of an interpretation of the IPCC concepts of 'confidence' and 'likelihood' that is compatible with the IPCC uncertainty guide's recommendations (and thus with the resulting practice of the IPCC authors in their communication of uncertainty). I argue that the ambiguity surrounding the concepts of 'likelihood' and 'confidence' has very serious and worrying implications for both the practice of the IPCC authors in their treatment of uncertainties and the quality of the information provided in the IPCC uncertainty report.

In Chapter 2, I argue that examining the history of the IPCC uncertainty framework, alongside the practice of the IPCC authors in their assessment of uncertainty, can shed some light on the conceptual problems in the current IPCC

uncertainty framework identified in Chapter 1. In particular, I argue that the persistent ambiguity in the relationship between 'confidence' and 'likelihood' can partly be traced back to the reasons behind the emergence of two uncertainty scales in the fourth assessment report. I show there were two distinct reasons for the emergence of these two uncertainty scales, that these two reasons are in clear tension with one another, and that the current AR5 Guide's recommendations are an unsuccessful attempt to deal with this tension. In an attempt to gain a better understanding of the IPCC concepts of 'likelihood' and 'confidence', I also have a close look at some of the methods (i.e. "multi-model ensemble methods") that are currently used by the IPCC authors to assess uncertainty in a finding. I conclude that these methods are not conceptually coherent methods for producing probabilities (independently of whether they are interpreted as objective or subjective probabilities) and hence for deciding what likelihood interval to assign to a finding. This fact, I argue, can give us some further insights into the nature of the reasons behind the emergence of two uncertainty scales in the IPCC uncertainty framework.

In the second part of my thesis, I review and assess several arguments that have been offered to defend the epistemic import of model robustness and relate this discussion to the context of climate model ensembles and climate scientists' current efforts to find an adequate measure of model independence. Part 2 consists of the following three chapters.

In Chapter 3, I discuss Weisberg's general characterization of robustness analysis and the role that he envisions for it in the discovery of "robust theorems". I argue that Weisberg's notion of low-level confirmation is unable to automatically confirm hypotheses that concern the actual world. Hence, I conclude that if low-level confirmation automatically confirms robust theorems, as Weisberg suggests, then robust theorems do not have to be hypotheses that are relevant to the explanation or prediction of real-world phenomena (as is usually assumed in the literature) for them to qualify as robust theorems.

In Chapter 4, I turn to various arguments that have been offered to support the idea that robustness analysis itself can confirm a robust theorem (which I

now interpret as the hypothesis that a causal structure of the model has a stable capacity to manifest a particular result of the model). I critically assess an argument put forward by Kuorikoski et al. (2010) for the epistemic import of model-based robustness analysis, an argument which I believe to be a formal expression of a widely held but ultimately misleading intuition: namely, the intuition that a model's conclusion is more likely to hold in the target system if several models lead to that conclusion because it would be a remarkable coincidence if that were not so. Kuorikoski et al. offer the best available defence of this intuition and that is why I believe it is important to rigorously assess it. I argue that, though Kuorikoski et al.'s argument relies on a weaker notion of probabilistic independence than unconditional independence, it cannot be sound. By relying on a different notion of independence (Fitelson's (2001) account of confirmational independence), I offer a revised, prima-facie more plausible argument. However, I show that this revised argument also relies on assumptions that are hardly ever plausible. Finally, I turn to Schupbach's (2018) recent account of robustness analysis as explanatory reasoning. I show that, although this account seems to fit well and in a straightforward manner with some empirical cases of robustness analysis, when one tries to apply Schupbach's account to model-based robustness analysis the picture is rather more complicated than Schupbach suggests, for its application relies on several non-trivial assumptions. Despite this, I argue that those assumptions may be reasonable in cases where the hypothesis we are interested in confirming through model-based RA is a 'robust theorem'. Hence my conclusion here is modestly positive: Schupbach's account could indeed be adequate (from a Bayesian perspective) for justifying why and determining when model-based RA should increase one's confidence in a 'robust theorem', and also for helping us understand the extent of that confirmation.

In Chapter 5, I review and critically assess some prominent arguments that have been offered by various philosophers (Lloyd, 2015; Parker, 2011; Justus, 2012; Winsberg, 2018) that could in principle (if not necessarily in practice) motivate the epistemic import of model robustness in the context of climate model

ensembles. I pay particular attention to Winsberg's (2018) recent and much celebrated suggestion that Schupbach's explanatory account of robustness analysis can finally shed light on the significance of the robustness of climate model ensembles' results. I argue that Schupbach's account is inapplicable whenever the models in an ensemble involve incompatible assumptions about a target system and the hypothesis we are interested in confirming concerns that target system. In light of this, I conclude that, despite Winsberg's emphatic suggestion, Schupbach's account cannot shed any light on the epistemic import of model robustness in climate science, because it is inapplicable. I then turn to consider what climate scientists have said about the epistemic import of model robustness. In particular, I focus on climate scientists' perennial search for an adequate measure of independence across climate models. I first review the various ways climate scientists have sought to define and measure independence across models, then consider the challenges each of these approaches faces. Finally, I argue that this arduous search is implicitly guided by an undefended and questionable assumption: that the more dissimilar models are from other models in an ensemble, the greater the confidence we should have in those models' consensus.

In the third and last part of this thesis, I explore the notion of the 'weight of evidence' typically associated with Keynes (1921). I argue that the Bayesian (or anyone who believes the role of probability in inductive inference is to quantify the degree of belief to assign to a hypothesis given the evidence) is bound to struggle with this notion, and suggest some lessons we might learn from the fact of this struggle. Although this discussion may appear far removed from any practical analysis of how the IPCC should characterize and communicate uncertainty in their findings, I show that a thorough understanding of the (problematic) nature of this notion is relevant to the assessment and evaluation of some recent proposals for a new IPCC uncertainty framework. Part 3 consists of the following two chapters.

In Chapter 6, I discuss in detail Keynes's (1921) often cited notion of the 'weight of evidence'. As we will see, Keynes understood the weight of evidence

in at least two different ways, and that it is ultimately impossible to directly measure Keynes's weight of evidence, however we choose to understand it. I then turn to the Bayesian's efforts to account for the weight of evidence. I argue that, contrary to what seems to be implicitly assumed in the literature, the Bayesian has not found an adequate way to account for the weight of evidence, and that it is unlikely they will ever do so, for several reasons. Finally, I suggest that the fact the Bayesian worries about the weight of evidence and yet struggles to provide an adequate response to those worries sheds light on the limitations of an epistemology that envisions the role of probability to be that of quantifying the degree of belief to assign to a hypothesis given the available evidence.

In Chapter 7, I critically assess three recent proposals for a new IPCC uncertainty framework that significantly depart from the current one. These proposals differ substantially from one another, and I believe these differences raise many philosophically interesting questions, some of which I attempt to address in this chapter. I first discuss Winsberg's (2018) proposal, according to which the likelihood metric should be used to communicate the range of credences that the IPCC authors accept it is rational to assign to a hypothesis in light of the available evidence, and the confidence metric should be used to communicate 'how likely their consensus regarding appropriate credences is going to remain fixed in the light of future developments' (ibid., 105). Amongst other things, I argue that Winsberg's interpretation of the confidence metric, under his own proposal, is unjustified. I then turn to Mach et al.'s (2017) proposal, which gets rid of the confidence metric and replaces it with qualitative terms for scientific understanding. I argue that Mach et al.'s proposal faces very similar conceptual problems to the current uncertainty framework (discussed in Part 1) and that it therefore does not constitute a considerable improvement. The last proposal I discuss is Bradley et al.'s (2017), according to which it should be possible to assign different likelihood levels qualified by different confidence levels to the same hypothesis, and that the IPCC authors should be encouraged to do so. I argue that the interpretation of confidence, under this proposal, is conceptually problematic. Finally, I offer my own tentative sketch for a new and better IPCC uncertainty framework, in particular, one that satisfies the two following

desiderata: 1) the framework's fundamental concepts should be clearly defined so that they can be used appropriately and consistently by the IPCC authors in the communication of uncertainty; 2) the use of the framework's fundamental concepts should help the IPCC authors produce findings that are interpretable, relevant and useful for the target audience.

Part I

An assessment of the IPCC conceptualization of uncertainty

Chapter 1

Some conceptual problems in the IPCC

1.1 Introduction

Studies of climate change are afflicted by deep uncertainty, the communication of which is made fraughter still by the studies' immediate policy implications. The world of policy-making has its demands: uncertain information should be communicated in a simple, consistent and relevant manner. It is thus vital to communicate this uncertainty in the most comprehensive, true-to-the-science and decision-relevant way, while making sure not to understate uncertainty. To address this, the fifth and latest assessment report (AR5) by the Intergovermental Panel on Climate Change (IPCC) makes extensive use of calibrated language to communicate uncertainty in its findings.[1] Below are some typical findings from the Summary for Policy Makers (IPCC 2013b) from the Working Group I

[1]The IPCC is an international body which synthesizes and communicates the current state of knowledge about climate change so as to 'provide a scientific basis for governments at all levels to develop climate related policies' (IPCC 2013a). Since its establishment in 1988 it has had five assessment cycles, each delivering an assessment report, and it is currently in its sixth assessment cycle with the Sixth Assessment Report (AR6) due to be completed by 2022. The Working Group I contribution to the Sixth Assessment Report (AR6) 'Climate Change 2021: The Physical Science Basis' was released at the very same time of this thesis' completion. This is why in this chapter, and the rest of this thesis, I solely focus on the communication of uncertainty by the AR5, rather than the AR6. However, although the IPCC uncertainty framework has gone through considerable revisions before each of the last three major assessment reports (the AR3, the AR4 and the AR5), as Janzwood (2020, 1656) reports 'the decision was made to not update the framework and implementation guidelines prior to the commencement of the Sixth Assessment Report (AR6) cycle'. Indeed, to the best of my knowledge, there have not been any significant changes in the communication of uncertainty by the IPCC from the AR5 to the AR6. Hence, I believe that a critical assessment of the AR5 reporting of uncertainty is equally relevant to that of the AR6.

(WG I) contribution to the AR5.[2] In some cases, both likelihood and confidence terms are used to communicate uncertainty in a finding:

1. Equilibrium climate sensitivity is *likely* in the range 1.5°C to 4.5°C (*high confidence*). (IPCC 2013b, 16; original emphasis)

2. Relative to the average from year 1850 to 1900, global surface temperature change by the end of the 21st century [. . . is] *unlikely* to exceed 2C for RCP2.6[3] (*medium confidence*). (ibid., 20; original emphasis)

In other cases, only a likelihood term is used:

3. It is *likely* that the frequency of heat waves has increased in large parts of Europe, Asia and Australia. (ibid., 5; original emphasis)

And in other cases still, only a confidence term is used:

4. There is *very high confidence* that the extent of Northern Hemisphere snow cover has decreased since the mid-20th century. (ibid., 9; original emphasis)

5. Annual CO_2 emissions from fossil fuel combustion and cement production were 8.3 [7.6 to 9.0] GtC12 yr–1 averaged over 2002–2011 (*high confidence*). (ibid., 12; original emphasis)

The presentation of these findings gives rise to several questions. What do these confidence and likelihood terms mean? Why is the IPCC using two metrics to communicate uncertainty in its findings and what is the relationship between them? Are they supposed to represent different types of uncertainty? If so, what types of uncertainty? Why is a likelihood assigned to some ranges (as in (1)) but not to others (as in (5))? What does it mean, in (2), to claim that warming is *unlikely* to exceed 2°C for RCP2.6 with *medium confidence*? If the IPCC has only

[2]There are three working groups, each responsible for a distinct part of an IPCC assessment report. WG I assesses the physical scientific basis of the climate system and climate change. WG II assesses the vulnerability of socio-economic and natural system to climate change, consequences and adaptation options. WG III assesses climate change mitigation methods.

[3]RCP2.6 is one of the four Representative Concentration Pathways (RCP) that have been adopted by the IPCC in the AR5 (together with RCP4.5, RCP6, and RCP8.5). The IPCC considers all four RCPs possible (but currently unverifiable) greenhouse gas concentration trajectories.

medium confidence then should one believe that warming is really *unlikely* to exceed 2°C for RCP2.6?

This chapter stems from an investigation into the above questions. Its aim is to offer a thorough diagnosis of some of the conceptual problems currently faced by the IPCC uncertainty framework as I believe that any successful attempt to revise and improve this framework will have to start from a clear understanding of the current conceptual problems, their implications for the IPCC authors' treatment of uncertainties, and the quality of the information provided in the AR5. Accordingly, this chapter (and the next) is an attempt to contribute to this first step.

The structure of this chapter is as follows. In Section 1.2, I will give a brief introduction of the current IPCC uncertainty framework. In Section 1.3, I will discuss the puzzling bifurcation between evidence and agreement in the characterization of confidence. In Section 1.4, I will argue that it is very unclear what types of uncertainty both the confidence and the likelihood metric are supposed to represent and that no matter what interpretation one gives to the IPCC concepts of 'confidence' and 'likelihood', none is compatible with some of the IPCC uncertainty guide's recommendations - and thus with the resulting practice of the IPCC authors in their communication of uncertainty. In Section 1.5, I will show that the ambiguity surrounding the concepts of 'likelihood' and 'confidence' has very serious and worrying implications for both the practice of the IPCC authors in their treatment of uncertainties and the quality of the information provided in the AR5. In Section 1.6, I will give a brief summary of the the conceptual problems in the IPCC uncertainty framework identified in this chapter and set forward the path for the next one.

1.2 The current IPCC uncertainty framework (for the AR5 and the AR6)

Any attempt to understand the interpretation of confidence and likelihood terms in the AR5 should of course start from an inspection of the AR5 uncertainty

guide (IPCC, 2010), henceforth referred to as the "Guide", to which I will now turn.[4]

The Guide provides both a confidence and a likelihood metric for experts to characterize uncertainty in their findings. The confidence metric is defined on a *qualitative* scale with five levels (“very low", “low", “medium", “high" and “very high"). The appropriate level of confidence depends on the evaluation of two independent dimensions: evidence and agreement. The evaluation of evidence can be low, medium or robust and depends on five criteria, namely the type[5], amount, quality, consistency and independence of the available evidence. The evaluation of agreement can be low, medium or high; what its evaluation depends on, however, is not specified in the Guide, but according to Mastrandrea et al. (2011, 678),[6] the degree of agreement is meant to express ‘a measure of the consensus across the scientific community on a given topic and not just across an author team’. The Guide specifies that although ‘increasing levels of evidence and degrees of agreement are correlated with increasing confidence’ the evidence and agreement dimensions are somewhat coarse grained, that is ‘for a given evidence and agreement statement, different confidence levels could be assigned’ (IPCC 2010, 3). Figure 1.1 shows the diagram provided by the Guide to illustrate the relationship between the evaluation of evidence and agreement and that of confidence.

Curiously however, despite what Figure 1.1 may suggest, the Guide further stresses that ‘confidence cannot necessarily be assigned for all combinations of evidence and agreement’ (ibid, 3); for some combinations the appropriate summary terms for the evaluation of evidence and agreement should be assigned instead.[7] The rules for when and when not to assign confidence are somewhat

[4]As I will discuss in Chapter 2, the IPCC uncertainty framework has gone through considerable revisions before each of the last three major assessment reports: the AR3, the AR4 and the AR5. However, as mentioned in Footnote 1, the AR5 uncertainty framework has not been updated prior to the commencement of the AR6, which will conclude in 2022. This is why the AR5 uncertainty framework which I will introduce in this section, can also be thought of as the *current* uncertainty framework for the AR6.

[5]The types of evidence included by the Guide are: ‘mechanistic understanding, theory, data, models, [and] expert judgment’ (IPCC 2010, 1).

[6]The commentary article by Mastrandrea et al. (2011) is an additional document provided by the IPCC to explain the AR5 uncertainty framework.

[7]Indeed, although in the Summary for Policy Makers, the WG I authors always assign an overall evaluation of confidence, the authors of WG II and III make frequent use of the summary terms. In this chapter, however, I will mainly focus on the practice of WG I

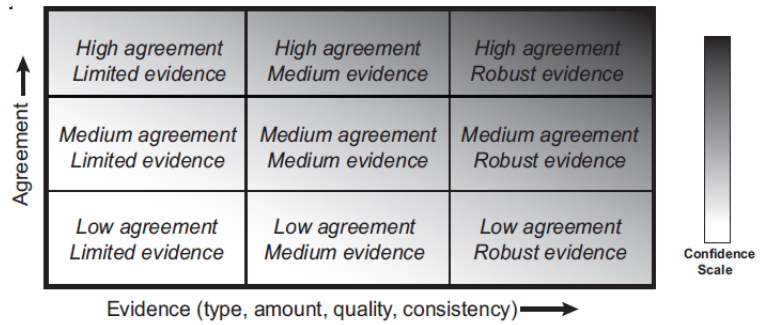

FIGURE 1.1: 'A depiction of evidence and agreement statements and their relationship to confidence. Confidence increases towards the top-right corner as suggested by the increasing strength of shading.' (ibid., 3)

ambiguous and can be summarised as follows (ibid, 2-3):

- confidence should be assigned in cases of high agreement and robust evidence and also, when possible,[8] in cases of with high agreement or robust evidence, but not both;
- confidence should not be assigned in cases of low agreement and limited evidence;
- the Guide does not specify whether or not confidence should be assigned in all the other cases.

The likelihood metric, on the other hand, is defined on a quantitative scale with seven levels: "Exceptionally unlikely", "very unlikely", unlikely", "about as likely as not", "likely", "very likely" and "virtually certain"; where each likelihood level corresponds to a probability interval[9] as shown in the table provided by the Guide (Figure 1.2). According to the Guide this metric is meant 'to express a probabilistic estimate of the occurrence of a single event or of an outcome' [. . .and it] may be based on statistical or modelling analyses, elicitation of expert views, or other quantitative analyses' (ibid., 3).

[8]The Guide does not specify why it would be possible to assign confidence in some cases, but not in others.

[9]The Guide specifies that each likelihood level 'can be considered to have "fuzzy" boundaries'. (ibid., 3)

Table 1. Likelihood Scale	
Term*	**Likelihood of the Outcome**
Virtually certain	99-100% probability
Very likely	90-100% probability
Likely	66-100% probability
About as likely as not	33 to 66% probability
Unlikely	0-33% probability
Very unlikely	0-10% probability
Exceptionally unlikely	0-1% probability

FIGURE 1.2: The likelihood metric

The Guide also has a few recommendations as to when and when not to use the likelihood metric. For a start, the Guide discourages authors from using the likelihood metric when 'probabilistic information' is not available:

> A likelihood or probability should be assigned for the occurrence of well-defined outcomes for which probabilistic information is available; (ibid., Annex B)

and it encourages to only use confidence in these cases, as in the following instruction for instance:

> If a range can be given for a variable, based on quantitative analysis or expert judgment: Assign likelihood or probability for that range when possible; *otherwise only assign confidence*. (ibid., 4, my emphasis)

These two recommendations may explain why there are cases where only a confidence term is used, as in findings (4) and (5) in Section 1.1. Importantly, the Guide also prohibits authors from using likelihood terms if the confidence level is not sufficiently high:

> [A likelihood] assignment should only be made when confidence is "high" or "very high," indicating a sufficient level of evidence and degree of agreement exist on which to base such a statement. (ibid., Annex B)

Although this recommendation is mostly followed by the AR5 authors as, for instance, in finding (1), where a likelihood term is used and the level of confidence is "high", this is not always the case. For instance, in finding (2), the level of confidence is "medium" (and therefore neither "high" nor "very high"), but a likelihood term is used nonetheless. In addition, the Guide does not always require an explicit mention of the level of confidence:

> A finding that includes a probabilistic measure of uncertainty does not require explicit mention of the level of confidence associated with that finding if the level of confidence is "high" or "very high". (ibid., 3)

This last recommendation may explain why there are cases where only a likelihood term is used such as in finding (3): perhaps the level of confidence associated with that finding is sufficiently high for the authors not to be required to explicitly mention it.

So, in brief, the Guide seems to outline the following process for evaluating and communicating uncertainties in findings. First and foremost, the authors are instructed to evaluate evidence and agreement for a finding. Next, if possible, the authors are instructed to assign confidence which will depend on the evaluation of evidence and agreement. Finally, if probabilistic information is available and confidence is sufficiently high, the authors are further instructed to assign likelihood (or a more precise presentation of probability). The diagram below (provided by Mastrandrea et al. (2011)) is a helpful illustration of this process (although for completeness I have added the red text).

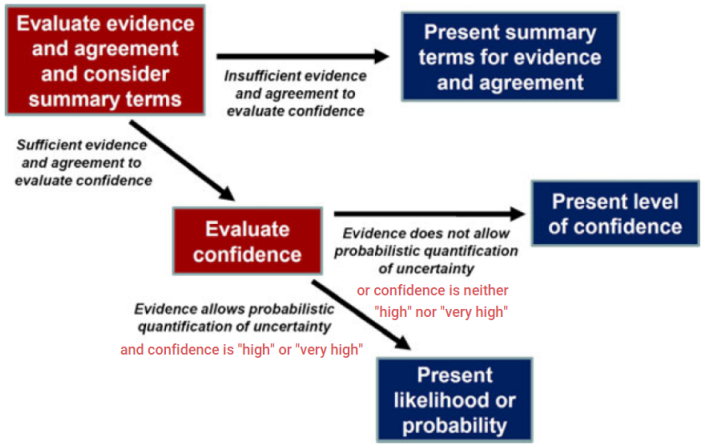

FIGURE 1.3: 'Process for Evaluating and Communicating the Degree of Certainty in Key Findings' (Mastrandrea et al. 2011, 679)

In this section, I tried to present an adequate summary of the AR5 uncertainty guide (IPCC, 2010). Despite my hope, however, this uncertainty framework does not provide a clear cut answer to the questions I asked in the introduction. It is still not clear to me what uncertainty confidence and likelihood terms are supposed to represent and what is the relationship between them. In the next two sections, I will discuss what I find the most perplexing aspects involved in the characterization of confidence and in its relationship with likelihood.

1.3 The perplexing bifurcation between evidence and agreement in the characterization of confidence

Given the definition of the evidence and the agreement metrics, there appears to be something rather problematic about the account of confidence provided: it seems clear that, differently from what this uncertainty framework seems to suggest, agreement and evidence so defined cannot be two independent dimensions; the level of agreement (understood as the level of scientific consensus on a given topic) surely *must* depend on the consistency, quality, amount and independence of the available evidence. Indeed, consider again Figure 1.1. Although the diagonal elements of the matrix are, somewhat plausible, the off-diagonal elements are, arguably, not. Take, for instance, a case with low agreement, but robust evidence. As Wuthrich (2017, 104) points out, it seems rather odd for there

to be low agreement (i.e. little scientific consensus on a given finding) when the evidence underpinning that finding is robust, 'given one makes the minimal assumption that scientists base their judgements on the available evidence'. Similarly, what are we to make of a case with high agreement, but limited evidence? It seems rather peculiar for there to be high scientific consensus on a given finding if the evidence underpinning that finding is limited: if it is not the robustness of the evidence that is driving the scientific consensus, then the presence of scientific consensus in this case should, arguably, be a cause of concern, rather than a reason to increase confidence.

Recall at this point that the Guide does stress that confidence cannot necessarily be assigned for all findings. In particular, for findings with high agreement or robust evidence but *not both*, it recommends authors to assign confidence only *when possible*. If not possible, the authors are instructed to assign the appropriate combination of summary terms for the evaluation of evidence and agreement instead. The Guide does not specify, however, *why* for findings with high agreement or robust evidence, but not both, it would be possible to assign confidence in some cases, but not in others, and why it should not be possible to assign confidence in the first place. This is rather puzzling given that Figure 1.1 does seem to specify confidence levels in these cases. Is the ambiguity surrounding the Guide's recommendations as to when and when not to assign confidence perhaps due to an undisclosed acknowledgement of the tension arising from the bifurcation of evidence and agreement in the characterization of confidence?

Finally and very curiously, Mastrandrea et al. (2011, 678) further state that:

> indicates, for example, the degree to which a finding follows from established, competing, or speculative scientific explanations. Agreement is not equivalent to consistency. Whether or not consistent evidence corresponds to a high degree of agreement is determined by other aspects of evidence such as its amount and quality. (agreement)

But this passage seems to articulate a rather different meaning of agreement altogether. It suggests that agreement is not to be understood as a measure of consensus in the scientific community, but rather as a measure of consistency

and other aspects of the available evidence. It is further suggested that the level of agreement for a given finding is in fact *determined* by the amount, quality and possibly other aspects of the available evidence. But given that these are criteria of the evidence metric, it is then very puzzling how the agreement dimension is supposed to be distinct from that of evidence in the first place.

Indeed, as Regh and Staley (2017, 132) point out, the very practice of the AR5 authors makes it very hard to sustain a consensus interpretation of the agreement metric:

> Should IPCC authors want to know the level of community acceptance, then polling the scientists in the relevant community might be more efficient, and in any case would appear to be the most important kind of evidence relevant to attributing community consensus.

But to their knowledge, and also mine, the AR5 does not report such a survey (nor does the Guide say that this should be done). One could argue that the IPCC authors might be able to *infer* the scientific consensus on an IPCC finding from the extent to which the set of relevant publications' conclusions agree with that finding (which according to Regh and Staley's examination of the AR5's practice is what agreement attributions actually track in most cases). But this is problematic for at least two reasons. For a start, as Regh and Staley (2017) point out, 'the IPCC authors must assume either that the teams involved in the cited research effectively exhaust the relevant community of scientists competent to judge the evidence for the claim, or that the acceptance of those teams in effect represents the best opinion of the relevant community', which as Regh and Staley remark, are not uncontroversial assumptions. Most importantly, however, if the IPCC authors were to *infer* the level of scientific consensus on a finding from a set of publications whose conclusions agree with it, then it seems that this inference would be determined by the evaluation of the evidence dimension, since that set of publications is surely part of the evidence underlying that finding! Hence, again it is very hard to see why the agreement dimension would be independent from that of evidence.

From the above discussion the following two conclusions seem to follow naturally:

1. if agreement is understood as a measure of consensus in the scientific community, then it is very unclear how evidence and agreement should be aggregated into an overall confidence judgement: the level of agreement *must* depend on the consistency, quality, amount and independence of the available evidence, hence the off-diagonal elements in Figure 1.1 make little, if any, sense;

2. if agreement is understood as a measure of consistency and other aspects of the available evidence, it is very unclear whether evidence and agreement are in fact distinct dimensions in the first place.

All this strongly suggests that there is a clear tension arising from the bifurcation of evidence and agreement in the characterization of confidence in the current AR5 uncertainty framework. Although I am not arguing that this apparent tension is unresolvable per se, I do think any attempt to resolve this tension would have to begin with giving an explicit and satisfactory answer to the following question: How are the evidence and the agreement metrics defined so that they are clearly independent from one another? For instance, here is an easy (but rather unsatisfactory) way out to resolve to some extent the tension arising from the bifurcation of evidence and agreement. Let's say the evidence metric's evaluation were to exclusively depend on considerations about the amount of evidence available (despite the fact that it is not clear at all how one should evaluate 'the amount' of available evidence in the first place) and not about other criteria (such as quality, consistency and independence), and the agreement' metric's evaluation were to depend on the level of consensus amongst the scientific community on the extent to which the available evidence supports a particular finding; this would seem to resolve the tension arising from the bifurcation of evidence and agreement. For instance the fact that there is a lot of evidence relevant to a particular finding and the fact there is little consensus on the extent to which that evidence supports that finding no longer seem incompatible: the

lack of consensus may be due, for instance, to the lack of consistency or independence amongst the different lines of evidence. However, this would evidently be an unsatisfactory solution. The extent to which the scientific community thinks the available evidence supports a particular finding should of course depend on considerations regarding the type, quality, consistency and independence of the available evidence; and by not making these considerations explicit, one would be simply accepting that the level of consensus in the scientific community is somehow representative (i.e. can be used as a proxy) of all such considerations. But this seems strange: one would be effectively evaluating the 'robustness' of the evidence underpinning a finding not by making explicit considerations regarding the available evidence, but by keeping track of the level of consensus amongst the scientific community regarding the extent to which the available evidence supports that finding. The epistemic justifications for this inference are dubious.

Despite the fact that it might be possible, in one way or another, to resolve the tension arising from the bifurcation between evidence and agreement in the characterization of confidence, it is doubtful there is an *epistemically warranted* way in which this can be done: *all* that should matter in the assessment of confidence underpinning a finding is the evaluation of the evidence underpinning that finding and nothing else. In other words the evaluation of the evidence underpinning a finding should *determine* the evaluation of confidence underpinning that finding. By this I am not at all trying to suggest that expert judgment does not or should not play a role in the assessment of confidence underpinning a finding. What I am arguing, however, is that that role should only enter in the very evaluation of the evidence underpinning that finding[10]. I am also not at all trying to suggest that experts always necessarily agree about the evaluation of the evidence with respect to a particular finding. For instance, as Douglas (2012, 152) notes, experts can very well look at the same evidence and come up with different explanations about why the evidence appears as it does; but

[10]Or potentially as a type of evidence. For example as Wuthrich (2017) points out one way in which expert judgment could potentially be considered as a type of evidence is 'when it concerns a piece of information for which there is no direct evidence (e.g., the tuning parameter values for a climate model)'.

the fact that there are distinct explanatory accounts consistent with the available evidence, is itself a fact that should be taken into account in the evaluation of the evidence underpinning a finding. This brings me to the evidence dimension. Recall that the evidence dimension is itself an aggregate of five criteria, namely the type, amount, quality, consistency and independence of the available evidence. Most of these criteria are not formally defined. But even if these criteria could be evaluated cogently when taken by themselves, it is not at all clear how they should be aggregated into an overall evaluation of the evidence dimension. For instance what "evidence level" should be assigned to a situation in which we have a high amount of high quality data that are independent but inconsistent? And what about a situation in which we have a high amount of low quality, independent and consistent data? It is clear that expert judgment will and, arguably, *should* play a role in the evaluation of the evidence dimension with respect to these criteria, but then the IPCC uncertainty framework should be *explicit* as to what that role is, rather than, as one might put it, 'hide' it away into a bewildering separate agreement dimension.

1.4 What types of uncertainty do confidence and likelihood represent?

As we have seen in section 1.2, the AR5 uncertainty guide provides not one but two metrics, likelihood and confidence, for the IPCC authors to assess and communicate uncertainty in their findings. The following question thus arises: what is the relationship between likelihood and confidence? That is to say, do likelihood and confidence represent different types of uncertainty, or not? And if so, what different types of uncertainty are these? Much as we might wish for an unambiguous answer to these questions, and despite what the IPCC authors seem to think,[11] the IPCC uncertainty guidance and the resulting practice of the IPCC authors make the above question a rather hard one to answer. Indeed,

[11] In the 'Annex A: Comparison of AR4 and AR5 Approaches' of the uncertainty guide it is asserted that compared to the previous uncertainty guide for the AR4, 'the AR5 guidance [...] is more explicit about the relationship and distinction between confidence and likelihood'. (IPCC 2010, Annex A)

I will argue that there is no interpretation of confidence and likelihood that is compatible with the Guide's recommendations.

Let me give a brief recap of what has been established so far. On the one hand, there is confidence, which is a *qualitative* scale used to synthesize 'the author teams' judgments about the validity of findings as determined through their evaluation of evidence and agreement' (IPCC 2010, 3). On the other hand, there is likelihood, which is a *quantitative* scale used 'to express a probabilistic estimate of the occurrence of a single event or of an outcome [... and it] may be based on statistical or modeling analyses, elicitation of expert views, or other quantitative analyses' (IPCC 2010, 3). Crucially, the Guide stresses that likelihood should only be assigned if confidence underpinning the finding is sufficiently high and if the available evidence allows a probabilistic quantification of uncertainty (IPCC 2010, 3-4 and Annex B; see also Figure 1.3).

So what to think of all this? According to Jones (2011, 736), a lead author on the third, fourth and fifth IPCC assessment report,

> The confidence-likelihood metrics of the uncertainty guidance form an epistemological ontological structure: confidence is epistemological and likelihood is ontological. The twinned basis for a key scientific finding combines ontological reasons—what the author team knows—and epistemological reasons—how confident are they in that knowledge— for a particular conclusion or set of conclusions [...]

Jones's terminology, here, is admittedly a little odd; usually one would say that epistemology, rather than ontology, is concerned with knowledge. But I think the only way to interpret this quote, that is what Jones really means to say here, is the following: whereas the likelihood metric is ontological, e.g. it expresses objective facts about the world, the confidence metric is epistemological e.g. it expresses the IPCC authors' confidence that their best theories offer a truthful reflection of those facts.

To understand Jones' interpretation of the likelihood and confidence metrics it is necessary, at this point, to make a distinction between *subjective* probabilities and *objective* probabilities. A fairly uncontroversial assumption in philosophy is

that the interpretations of probability fall into two broad families. *Subjective* interpretations view probabilities as dependent on the mental states of individual agents. These probabilities are referred to as 'subjective probabilities', 'credences' or 'degrees of belief'. Whereas *objective* interpretations view probabilities as features of the mind-independent world. These probabilities are usually referred to as 'objective probabilities' or 'chances'. For the purpose of this discussion I will assume that chances are in some way connected to frequencies of an outcome and that it is possible to make true chance statements about a system that obeys deterministic laws, that is I will assume that chances and determinism are compatible.[12] But then if likelihood is ontological and hence expresses an objective fact about the world, as Jones claims, the right interpretation of probability, as far as the likelihood metric is concerned, must be an *objective* one. In other words, if likelihood is ontological, then it must be used to express the *chance* of a single event or an outcome.

Under Jones's interpretation then, confidence and likelihood are understood as representing different types of uncertainty: whereas the likelihood metric expresses the objective fact that the chance of the occurrence of an event or an outcome is in a particular interval, the confidence metric expresses the IPPC authors' confidence in the truth of this fact. Consider, for instance, the following finding from Section 1.1:

1. Equilibrium climate sensitivity is *likely* in the range 1.5°C to 4.5°C (*high confidence*) [. . .]; (IPCC 2013b, 16)

Under the above interpretation, it is possible to understand this statement; what it would mean is as follows: the world is such that the objective probability for the Equilibrium Climate Sensitivity to lie in the range 1.5°C to 4.5°C is included in the interval [0.66, 1]; and the IPCC authors have high confidence in this finding. Under Jones' interpretation, what makes it possible to have two distinct uncertainty scales to characterize uncertainty in a finding is really a distinction between credences and chances: in this case the IPCC authors have high credences in the objective fact that the chance is in the interval [0.66, 1].

[12] See Frigg (2014) for a survey of compatibilist and incompatibilist positions.

However, given the nature of climate science, there are reasons to suspect that the idea that likelihood expresses something objective in the world is implausible. For although the IPCC does often rely on seemingly 'objective' methods (e.g. perturbed-physics ensemble methods and multi-model ensemble methods) to calculate the 'probability' of an event or an outcome (which will determine the likelihood assigned to that event or outcome), as Winsberg points out, these methods 'are *objective* only in the sense that they are independent of the degrees of belief of any particular expert, and they are calculated mechanically' (Winsberg 2018, 96). So even (or rather, especially)[13] in light of these 'objective' methods, it is very unclear how one should interpret probabilities in the IPCC report, which means that it is impossible to tell what kind of uncertainty the likelihood metric is *actually* supposed to represent.[14] It is also worth pointing out that, although the Guide itself does not give a general definition of probability, some of the Guide's remarks make it clear that at least in some cases, probability will have a subjective interpretation, such as this one: 'Where practical, formal expert elicitation procedures should be used to obtain *subjective* probabilities for key results' (IPCC 2010, Annex B, my emphasis).

Perhaps we should not, then, think of the probabilities in the IPCC assessment report as *objective* after all. But this is troubling, for if likelihood is not ontological as Jones claims it is, then it is no longer obvious how one should interpret the above statement. So the question is: is there another way to interpret it? Could we perhaps think of the interval [0.66, 1] as imprecise *subjective* probabilities, representing the IPCC authors' degrees of belief that the Equilibrium Climate Sensitivity lies in the range 1.5°C to 4.5°C? But in that case, what should we think of the qualifier "high confidence"? Winsberg (2018) claims one possible way to understand measures of confidence might be 'as a kind of second-order probability'; that is to say, the high confidence in the imprecise credence [0.66, 1] above

[13] Indeed, there are several reasons to suspect these methods are not conceptually coherent. I will discuss these reasons in detail in section 2.5 (see also Winsberg (2018, 97-100).

[14] I am certainly not alone in my perplexity as to how the IPCC interprets the concept of likelihood; see for instance Wuthrich (2017), Aven (2018) and Janzhood (2020), who after interviewing several IPCC authors, concludes that 'most authors agreed that the confidence/likelihood distinction was confusing' (ibid. 1670).

> [...] *is a bit like a high degree of belief that the credence will be resilient in the face of future evidence*- assessed by looking at the variety of evidence supporting the credence, and the degree of agreement among those sources supporting the credence. But given the general murkiness of second-order probabilities in general, the lack of an obvious set of decision rules to apply them, and the difficulties that would be involved in interpreting such probabilities in this specific case, [...] it is wise of the IPCC to refrain from using the expression 'probability' for its second-order characterization, and to limit itself to qualitative characterization of confidence. (Winsberg 2018, 105, emphasis in the original)

In my view, to speak of wisdom here requires rather too much indulgence. First, there is no good evidence that should lead one to believe this is what the IPCC has in mind. That is, there is no evidence to suggest the IPCC authors might have such an epistemologically sophisticated thing in mind. Second, Winsberg proposes here a particularly advanced epistemological interpretation, one that certainly invites critical assessment.[15] But in this chapter, I will leave that assessment to one side; I will instead follow each of these two *distinct* candidates for possible interpretations of likelihood and confidence (Jones's and Winsberg's) and see where they lead us, i.e. whether either is compatible with the Guide's

[15] In particular, and despite Winsberg's light-hearted attitude, it is not *at all* obvious how one would go about assessing the resiliency of one's beliefs in the face of future evidence; nor why, for instance, the amount of available evidence (which is one of the factors supposed to affect the evaluation of confidence - see Figure 1.1) would help with this assessment. As I will discuss in Chapter 7, I suspect Winsberg here is inspired by the Bayesian's perennial attempt to account for Keynes's notion of 'the weight of evidence', a notion that is not reflected in an agent's credence in a hypothesis. The Bayesian's standard reply is the following: true, 'the weight of evidence' is not reflected in an agent's credence in a hypothesis, but it is reflected in the *resiliency* of the agent's credence in that hypothesis (see e.g. Skyrms (1977), Joyce (2005)). However, the examples the Bayesian relies on to convince us of this are always, to the best of my knowledge, ones in which an agent's credence in a hypothesis is mediated by her beliefs about the hypothesis's objective chances. Hence, those cases do not remotely support the idea that the weight of evidence can manifest itself in the resiliency of an agent's credence in a hypothesis in cases where that credence is *not* mediated by her beliefs about the hypothesis' objective chances. This is, in essence, why I think Winsberg's account of confidence is problematic and ill-conceived; I will give a thorough assessment of Winsberg's account in Chapter 7 (Section 2). In any case, for the purpose of what I am arguing here, it can just be assumed that the level of confidence is merely an *additional* evaluation of the evidence (somehow based on factors such as amount, quality, etc., as shown in Figure 1.1) that was used by the IPCC authors to determine the range of credences that one ought to have in a hypothesis (Aven (2018, 290) calls this the 'strength of the knowledge' supporting a subjective probability judgment).

recommendations and the practice of the IPCC authors. I will argue that none of these possible interpretations fits the bill.

1.4.1 No interpretation fits the bill

Recall that likelihood is presented by the Guide as a subsequent option for characterizing uncertainty, following the evaluation of confidence. In other words, the authors are instructed to always *first* evaluate confidence, and only once this is done, *if* the evidence allows a probabilistic quantification of uncertainty and *if* confidence is sufficiently high, the Guide recommends authors to *also* assign likelihood (see Figure 1.3). I will argue that these recommendations are incompatible with either of the two interpretations of confidence and likelihood discussed in the previous section.[16]

Consider first Winsberg's interpretation. Here, confidence is a judgment about the *resiliency* of the IPCC's author's *credences* in a finding in the face of future evidence. In finding (1) (from Section 1.1.), for instance, the finding is: 'equilibrium climate sensitivity lies in the range 1.5°C to 4.5°C', the IPCC authors have credences $[0.66, 1]$ in this finding, and "high confidence" is a statement about the resiliency of *those credences* in the face of future evidence. But then, in cases where confidence is not sufficiently high, and hence likelihood is not assigned in accordance with the Guide's recommendations, one may very well ask: confidence is a statement about the resiliency of *what credences*? To clarify my concerns here, consider this particular instruction by the Guide:

> [If] a range can be given for a variable, based on quantitative analysis or expert judgment: Assign likelihood or probability for that range when possible; *otherwise only assign confidence*. (IPCC 2010, 4, my emphasis)

[16] I will exclusively focus on the Guide's recommendation to assign likelihood only if confidence is sufficiently high. However, the reasons that I give for the incompatibility between this recommendation and Winsberg's interpretation of confidence and likelihood also apply to the Guide's recommendation to assign likelihood only if the evidence allows a probabilistic quantification of uncertainty. In contrast, this latter recommendation *may* be compatible with Jones's interpretation of confidence and likelihood in so far as if there are no chances to report then the likelihood metric has no role to play.

Let us then consider a case where a range can be given for a variable but confidence is not sufficiently high. As instructed by the Guide, the authors only assign confidence. However, if confidence *had been* sufficiently high then the authors might have *also* reported likelihood for that range. But then under the above interpretation of confidence and likelihood it seems impossible to understand how to interpret such a finding: the given range could in principle have *any* likelihood assigned to it, and we are not told which! Indeed, below are two instances of such a finding:

> The release of CO2 or CH4 to the atmosphere from thawing permafrost carbon stocks over the 21st century is assessed to be in the range of 50 to 250 GtC for RCP8.5 (*low confidence*) (IPCC 2013b, 27; original emphasis);
>
> Annual CO2 emissions from fossil fuel combustion and cement production were 8.3 [7.6 to 9.0] GtC12 yr–1 averaged over 2002–2011 (*high confidence*) (IPCC 2013b, 12; original emphasis).

In these cases, the authors do not assign a likelihood to the range 50 to 250 GtC in the first finding nor to the range 7.6 to 9.0 GtC12 yr–1 in the second finding and instead only assign a confidence statement. So how should one interpret these findings? Do the IPCC authors think these ranges are *likely*, *very likely* or perhaps *virtually certain*? Without this information it seems impossible to understand how to interpret these statements. As another instance, consider the following finding, in which the assessed range's endpoints are assigned different confidence levels:

> There is *high confidence* that sustained warming greater than some threshold would lead to the near-complete loss of the Greenland ice sheet over a millennium or more [...] the threshold is greater than about 1°C (*low confidence*) but less than about 4°C (*medium confidence*) [...]. (IPCC 2013b, 29; original emphasis)

Again no likelihood is assigned in this case, so it is very unclear how one should interpret this finding. Do the IPCC authors think the threshold is "likely" greater

than about 1°C, "extremely likely" greater than about 1°C or . . .? Do the IPCC authors think the threshold is *likely* less than about 4°C, *extremely likely* greater than about 1°C or . . .? Again without an answer to these questions and under Winsberg's interpretation of confidence and likelihood, it seems impossible to understand the meaning of this statement. Winsberg's interpretation thus appears incompatible with the frequent IPCC practice of only reporting confidence levels in a finding: if confidence is really a statement about the resiliency of the IPCC authors' credences in a finding then we need to be told what those credences are![17]

Consider now Jones's interpretation. Jones' interpretation is also not compatible with the recommendation to use the likelihood metric only in cases of sufficiently high confidence. If likelihood expresses the objective fact that an event or an outcome has a particular *chance*, then there seems no plausible reason for discouraging authors from communicating this fact even when confidence in the truth of this fact is not high. For instance, suppose that I have *medium confidence* in the claim that a coin is very biased towards Heads (e.g. it is *very likely* to land heads). If in this case I can't use the likelihood metric, what should I say about this coin? If I do not report likelihood and just claim that I have medium confidence in the coin landing heads, I'm omitting essential information: the *medium confidence* was a statement about the finding 'the coin is *likely* to land heads'. So if I just report confidence I would be failing to report the finding I actually have medium confidence in!

To recapitulate, both Jones' interpretation, where 'likelihood' is ontological and 'confidence' is epistemological and Winsberg's interpretation, where both 'likelihood' and 'confidence' are epistemological, are incompatible with the Guide's recommendation to assign likelihood only if 'confidence' is sufficiently high. But I can see no other plausible interpretation of the confidence and likelihood metric compatible with this recommendation; I am tempted to conclude that the problem is not my lack of imagination, but rather that uncertainty is not adequately conceptualized.

[17] The example in the previous quote is not an isolated instance nor is it cherry picked. It is one of the many instances in which the IPCC only assigns confidence to a finding.

1.5 Troubling implications

In this section, I will argue that the lack of a conceptually coherent interpretation of the concepts of 'confidence' and 'likelihood' in the current IPCC uncertainty framework has serious and worrying implications for both the practice of the IPCC authors in their treatment of uncertainties and the quality of the information provided in the AR5.

1.5.1 The lack of transparency behind the interaction between confidence and likelihood levels

There is a certain practice of the IPCC authors that may have something to do with the Guide's instructions to report likelihood only if the confidence underpinning those likelihood assignments is sufficiently high. Mach's (2017) analysis of the AR5's author's reporting of uncertainty reveals that authors often subjectively adjusted (i.e. downgraded) the likelihood given by the formal analyses; they did this to implicitly account for unquantified uncertainties (e.g. structural uncertainties in models), and thereby *raise* the level of confidence associated with that likelihood assignment. As Mach et al. (2017, 8) remark, this practice, although common, was rarely made explicit and transparent. An instance of this practice is the following:

> Increase of global mean surface temperatures for 2081–2100 relative to 1986–2005 is projected to *likely* be in the ranges derived from the concentration-driven CMIP5 model simulations, that is, 0.3°C to 1.7°C (RCP2.6) [. . .] (*Very high confidence*). (IPCC 2013b, 20; original emphasis)

In this case the 5-95% model ranges, which according to the Guide's calibrated language for likelihood terms correspond to *very likely* ranges, were instead interpreted as merely *likely* ranges in order to account for a not high enough confidence in the validity of the models.[18] As Frigg et al. (2015, 973) explain, this is

[18] According to Mach et al. (2017, 8) 'the 5-95% model ranges could conceivably represent very-likely ranges if the models were judged to include all relevant uncertainties; instead, using 5–95% ranges as likely ranges implies a subjective adjustment based on confidence in the validity of

a case where

> the ensemble information is used but supplemented with expert judgement about the chance that models are misinformative. In effect, 24% of the probability mass has been reassigned in an undetermined manner, which we might interpret as a probability of approximately up to one-in-four that something occurs which the models are incapable of simulating.

I believe this practice may have something to do with the Guide's instructions to report likelihood only if the confidence underpinning those likelihood assignments is sufficiently high: given that confidence in the validity of the underlying models (or rather in the assumptions that would justify interpreting the 5–95% model ranges as objectively 'very likely' - see footnote 18) was not sufficiently high to report likelihood terms, the IPCC authors may have chosen to downgrade the likelihood assignment so as to *increase* the confidence underpinning the conclusion, and thereby be in line with the Guide's directions.

But regardless of the Guide's instructions' responsibility for this practice, there seems to be a clear interaction between confidence and likelihood levels. This, for a start, seems to be in stark contradiction with the Guide's recommendation to assign confidence *prior* to assigning likelihood (given the fact that the AR5 authors can *upgrade* confidence by *downgrading* likelihood!); which, in my view, makes any attempt to understand what kind of uncertainty the confidence and likelihood metrics are meant to represent in the AR5 doomed from the outset.[19] Crucially, however, the IPCC uncertainty Guide says nothing about the

the underlying models.' However, despite what Mach et al. suggest it is very unclear what assumptions would have to hold for the 5-95% model ranges to conceivably represent *very likely* ranges. What Mach et al. (and I) call the 5-95% model ranges are in fact the 5-95% ranges of a normal distribution with the mean and standard deviation of the model ensemble's projections. *If* this normal distribution were to represent the actual uncertainty about future conditions, then of course the 5-95% ranges from this distribution should be considered to be *very likely* in accordance to the Guide's calibrated language for likelihood terms. However, it is extremely unclear what assumptions would in fact have to hold for this distribution to conceivably represent the actual uncertainty about the future conditions. Winsberg (2018, 97-102) discusses some, if not all, of those assumptions and concludes that they can never be plausibly satisfied. Hence, given this, the very idea that there are some conditions under which the ranges derived by these methods could be straightforwardly interpreted as *very likely* ranges is in my view misguided.

[19] In particular, consider Winsberg's interpretation: why should the downgrading of likelihood (i.e. choosing to report a broader likelihood interval), increase one's degrees of belief that a likelihood assignment will be resilient in the face of future evidence? I will come back to this question in Chapter 7, where I will critically assess Winsberg's proposal.

degree of this interaction (i.e. to what degree should likelihood be downgraded as confidence is upgraded). This is worrying for although the downgrading of likelihood in these cases was evidently intended to account for sources of uncertainty not adequately addressed in the formal analyses, it is very unclear to what extent they were accounted for.[20] In other words, it is often hard to tell on what basis the IPCC authors come to the conclusions one finds in the AR5 uncertainty report. Why, for instance, in the case discussed above, did the IPCC authors downgrade likelihood from *very likely* to *likely* rather than *more likely than not* in the process of upgrading confidence? What is the reasoning behind this conclusion? And is it good reasoning? What would good reasoning consist in? I think the apparent difficulty in answering the last question, combined with the lack of transparency of the IPCC's practice of downgrading likelihood thereby upgrading confidence, can only raise suspicions that this issue is taken much less seriously than it should be.

There is, I believe, an additional reason to be wary of the AR5 authors' common practice of downgrading likelihood, thereby upgrading confidence. If the authors lacked *very high confidence* in the validity of the models to assign a very likely assignment to the 5-95% model ranges, then why did they insist on reporting those very same ranges (but with a lower likelihood assignment)? That is, intuitively another way to upgrade confidence in the finding would have been to report a wider range instead, one that the authors considered *very likely* rather than just *likely*. So not only is the reasoning behind downgrading likelihood, thereby upgrading confidence, non-transparent and unclear, but what is also unclear is why the authors are choosing to downgrade likelihood in the first place when other options seem to be available.[21]

[20] See Parker and Risbey (2015) for some kind of considerations that would seem appropriate when it comes to downgrading likelihood.

[21] In my view, one reason to be very wary of this practice of downgrading likelihood rather than reporting a wider range is that, despite the downgrading of likelihood in this case, it is evident that the most salient feature of this finding is the range 0.3°to 1.7°; and given the substantial uncertainty in this range it seems to me highly misleading to draw so much attention to it. I will discuss this point further in Section 2.5.

1.5.2 Value judgments and non-interpretable findings

As it tuns out, and as already mentioned in Section 1.2, despite the Guide's instructions to report likelihood only if confidence is *high* or *very high*, there are in fact several cases where the AR5 authors do not follow these instructions; see, for instance, the following finding:

> Equilibrium climate sensitivity is [...] *very unlikely* greater than 6°C (*medium confidence*) (IPCC 2013b, 16; original emphasis)

But this is very puzzling. Why are the AR5 authors choosing to upgrade confidence in some cases (by reporting a broader likelihood interval)[22], but not in others? What is determining the IPCC authors' choice of the confidence level at which to communicate a particular finding? Indeed, if confidence levels can interact with likelihood assignments, as the practice of the AR5 authors suggests, then as some have noted (Bradley et al. (2017), Winsberg (2018)), the choice of reporting findings at a particular confidence level seems to involve a *substantial* (non-epistemic) value judgment; But if this is so, is it really up to the IPCC authors to make that value judgment? According to Winsberg, 'it seems clear that at least sometimes it is considerations of the likely applications of an uncertainty report that guide the choice between a wider and more confident report and a narrower and somewhat less confident report' (Winsberg 2018, 149), but then such considerations should be made explicit and open to scrutinization – which currently they are, quite evidently, not.[23]

Not only does the choice the of reporting probabilistic findings at a particular confidence level seems to involve a substantial non-epistemic value judgment, but it also makes the IPCC findings very hard to interpret. To see why this is, consider what decision making under uncertainty is all about. Standard decision theory recommends the agent perform the action that maximizes the expected utility relative to the probability of the states of the world and the

[22]The *very likely* range and the *likely* range correspond to an interval probability assignment of (0.9,1) and (0.66,1) respectively. Hence downgrading likelihood from *very likely* to *likely* is choosing to report a broader likelihood interval.

[23]I will come back to the question of whether, under Winsberg's own interpretation of 'confidence' and 'likelihood', value judgments should affect the choice of what likelihood and confidence levels to assign in Chapter 7.

utilities of the possible consequences of the actions. However, several scholars have noted that there are cases where an agent might hold imprecise probabilities rather than precise probabilities (see, for instance, Levi 1985; Bradley 2009; Gilboa et al. 2009; Joyce 2010). And in light of this view, several decision rules have been proposed to also deal with imprecise probabilities. Decision theory then seems to deal relatively well when faced with both precise and imprecise probabilities. However, the probabilistic findings in the IPCC are not expressed merely in terms of precise probabilities, nor imprecise probabilities! Rather they are mostly expressed in terms of imprecise probabilities, *qualified* by qualitative confidence judgements and this information seems really rather hard to integrate sensibly into an account of decision making. So the concern is the following: What role, if any, can these qualitative confidence judgments play in decision making? This is a legitimate concern. The main, if not only, reason an institution such as the IPCC is in place is to give relevant and useful information regarding the state of knowledge in studies of climate change to agents that will ultimately want to make decisions based on this information. If it is not at all clear how one should interpret this information so as to make rational decisions based on it, then the project seems to have (at least partly) failed.

1.6 Taking Stock

In this Chapter, I have discussed what I consider to be some serious conceptual problems in the current IPCC uncertainty framework. These were:

1. *The puzzling bifurcation of 'evidence' and 'agreement' in the characterization of confidence*;

   In Section 1.3, I argued that:

   - if agreement is understood as a measure of consensus in the scientific community, then it is very unclear how evidence and agreement should be aggregated into an overall confidence judgement: the level of agreement *must* depend on the consistency, quality, amount and

independence of the available evidence, hence the off-diagonal elements in Figure 1.1 make little, if any, sense;

- if agreement is understood as a measure of consistency and other aspects of the available evidence, it is very unclear whether evidence and agreement are in fact distinct dimensions in the first place.

I further argued that, although the tension arising from the bifurcation between 'agreement' and 'evidence' in the characterization of confidence is not unresolvable per se, any attempt to resolve this tension is likely to be inadequate from an epistemological point of view.

2. *The extremely ambiguous relationship between confidence and likelihood*;

   In Section 1.4, I argued that the relationship between 'confidence' and 'likelihood' is extremely ambiguous in so far as there seems to be no possible interpretation of confidence and likelihood that is compatible with the IPCC Guide's recommendation to only assign 'likelihood' if 'confidence' is sufficiently high.

In Section 1.5, I argued that the lack of a conceptually coherent interpretation of the concepts of 'confidence' and 'likelihood' has serious and worrying implications for both the practice of the IPCC authors in their treatment of uncertainties and the quality of the information provided in the AR5. In particular, I argued that one should be wary of the AR5 authors' frequent practice of subjectively *downgrading* the 'likelihood' given by the evidence, thereby *raising* the level of confidence associated with that likelihood assignment. This is due to two reasons:

1. The first has to do with *the lack of transparency* behind this practice. The Guide recommends authors to assign confidence based on the evaluation of evidence and agreement *prior* to assigning likelihood and hence has nothing to say about the degree to which likelihood should be downgraded as confidence is upgraded. Hence, although the IPCC authors' downgrading of likelihood is evidently intended to account for sources of uncertainty not adequately addressed in the formal analyses, it is very

unclear to what extent they are accounted for, as the reasoning behind this practice is very unclear. What is also not clear is why the authors are choosing to downgrade likelihood, thereby upgrading confidence, given that there seems other ways to upgrade confidence (e.g. to report a wider range instead);

2. The second has to do with *the lack of ubiquity* of this practice. Indeed there are cases where likelihood is reported with levels of confidence that are neither *high* nor *very high*. But then it is not clear what determines the IPCC authors' choice of the confidence level at which to communicate a particular probabilistic finding. Indeed, if confidence levels can interact with likelihood assignments, as the practice of the AR5 authors suggests, the choice of reporting findings at a particular confidence level seems to involve a substantial (non-epistemic) value judgment. But if this is so, then these value judgments should be made explicit and open to scrutinization, and currently they are not. I have further argued that the lack of a clear role for these qualitative confidence judgments to play in decision making, means it is not at all clear how one should interpret likelihood assignments qualified by confidence judgments so to ultimately make decisions based on them.

I hope that I have convinced the reader that these are not merely philosophical problems that can be left for a rainy day; these are serious conceptual problems in the conceptualization of uncertainty that have implications for both the practice of the IPCC authors in their treatment of uncertainties and the quality of the information provided in the AR5. In the next chapter, I will explore the extent to which the history of the IPCC uncertainty framework can shed light on the nature of these conceptual problems.

Chapter 2

A genealogy of 'confidence' and 'likelihood' and what we can learn from it

2.1 Introduction

Although the Second Assessment Report (AR2) included some discussion of the need for a framework that could be consistently applied for the communication of uncertainty across the three working groups (McBean et al., 1996), it was only as part of the Third Assessment Report (AR3) that uncertainty guidance (Moss and Schneider, 2000) was developed in an attempt to meet that challenge and encourage a more transparent and consistent treatment of uncertainty. An important step in this direction was made with the presentation of a calibrated uncertainty language to be used consistently amongst the working groups. However, since its initial presentation in Moss and Schneider's (2000) uncertainty guide, the IPCC's uncertainty language has been subject to considerable change over the course of the assessments; I believe an investigation into the nature of this change and the reasons for it can offer some important insights with respect to the issues identified in Chapter 1.

The structure of this chapter is as follows. In Section 2.2, I will present the uncertainty framework for the AR3 (Moss and Schneider, 2000), which included a *single* uncertainty scale (the confidence scale). I will then point to some of the

resemblances and discrepancies between this framework and the AR5 framework which I believe can shed light on the problematic bifurcation of evidence and agreement in the characterization of confidence discussed in Chapter 1. Finally, I will discuss WG I's decision to introduce, in the AR3 itself, an *additional* uncertainty scale (the likelihood scale) alongside the confidence scale. In Section 2.3, I will present the substantially revised uncertainty framework for the AR4 (IPCC, 2005), now including both a confidence and a likelihood scale. I will point to some problematic aspects of this uncertainty framework, aspects which motivated further revisions to the IPCC uncertainty framework. These revisions, whose main goal was, apparently, to clarify the 'distinction and transition' (IPCC, 2010) between the confidence and the likelihood scales, gave rise to the current uncertainty framework for the AR5. In Section 2.4, I will argue that the persistent ambiguity of the relationship between confidence and likelihood in the AR5 uncertainty framework can partly be traced back to the reasons behind the emergence of two uncertainty scales in the IPCC uncertainty framework. In particular, I will argue that there is a clear tension arising from these distinct reasons and that the current guide's recommendation to assign likelihood only if confidence is sufficiently high is an unsuccessful attempt to deal with this tension. In Section 2.5, in an attempt to gain a better understanding of the AR5 concepts of *likelihood* and *confidence*, I will have a close look at some of the methods (i.e. "multi-model ensemble methods") that are currently used by the IPCC authors to assess uncertainty in a finding. I will conclude that these methods are not conceptually coherent methods for producing probabilities (of any kind) and hence for deciding what likelihood interval to assign to a finding. In Section 2.6, I will take stock of what both the history of the IPCC uncertainty framework and the practice of the IPCC authors in their assessment of uncertainty can teach us about the conceptual problems in the current IPCC uncertainty framework identified in Chapter 1.

2.2 The first IPCC uncertainty framework (for the AR3)

The AR3 uncertainty guide (Moss and Schneider, 2000) presented a single quantitative confidence scale, and encouraged all working groups to use this scale to characterize the state of their knowledge underlying a finding, asserting that '[w]ithout such a discrete quantitative scale, there is strong experimental evidence that the same uncertainty words often have very different meanings for different people in different circumstances' (ibid., 44). The scale had five levels, ranging from "Very Low Confidence" to "Very High Confidence", where each level corresponded to a given probability interval as illustrated in Figure 2.1.

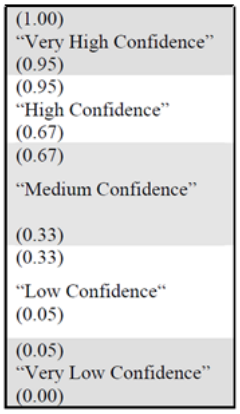

FIGURE 2.1: 'Scale for Assessing State of Knowledge' (ibid., 44)

The AR3 guide further specified that the appropriate interpretation of probability in most cases would be a *subjective* one: more specifically one according to which 'the probability of an event is the degree of belief that exists among lead authors and reviewers that the event will occur, given the observations, modeling results, and theory currently available' (ibid., 36).

Although, this was the only uncertainty scale provided by the AR3 guide for the IPCC authors to communicate uncertainty in their findings, the AR3 guide also remarked that in light of previous comments on earlier drafts, it was expected that 'some may be uncomfortable with having only one option' (ibid.,44) to characterize uncertainty. However, the nature of this discomfort was not explained. That is, although the AR3 guide implied that not everyone would be comfortable with only having one option to characterize uncertainty, it wasn't

clarified why some may feel more comfortable than others with only having one option. In any case, in light of this potential discomfort, the AR3 guide further proposed an additional set of qualitative uncertainty terms, to supplement the above quantitative confidence scale and enable the authors to explain the reasoning behind their level of confidence in a particular finding. The guide stressed, however, that these supplementary terms were indeed to be treated as *supplementary*, since 'these qualitative terms do not always map well onto a quantitative scale, increasing the likelihood of inconsistent usage' (ibid., 44). They were defined as follows (ibid., 45):

- *Well-established*: 'models incorporate known processes; observations largely consistent with models for important variables; or multiple lines of evidence support the finding';
- *Established but Incomplete*: 'models incorporate most known processes, although some parameterizations may not be well tested; observations are somewhat consistent with theoretical or model results but incomplete; current empirical estimates are well founded, but the possibility of changes in governing processes over time is considerable; or only one or a few lines of evidence support the finding';
- *Competing Explanations*: 'different model representations account for different aspects of observations or evidence, or incorporate different aspects of key processes, leading to competing explanations';
- *Speculative*: 'conceptually plausible ideas that haven't received much attention in the literature or that are laced with difficult to reduce uncertainties or have few available observational tests'.

These qualitative uncertainty terms were further situated in the table below, suggesting that one should think of them as jointly describing the 'amount of evidence' and the 'level of agreement/consensus' underlying a particular finding.

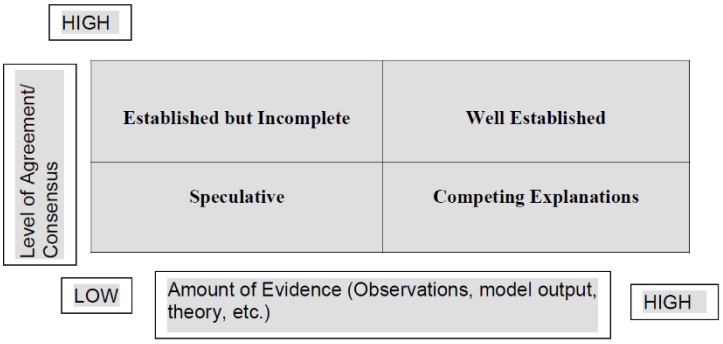

FIGURE 2.2: 'Supplemental Qualitative Uncertainty Terms' (ibid., 45)

The difference between the uncertainty language presented by this uncertainty guide and the one presented by the AR5 uncertainty guide (IPCC 2005) discussed in Chapter 1, is substantial. Arguably, the most striking difference is that this uncertainty guide presented only a *single* scale, i.e. the confidence scale, for the authors to characterize uncertainty in their findings (although a set of qualitative uncertainty terms was also presented, these were only meant to be used as *supplementary* terms). This is in stark contrast to the AR5 framework, which, as extensively discussed in Chapter 1, consists of two distinct uncertainty scales (i.e. 'confidence' and 'likelihood'). Another significant difference is that in this uncertainty guide the confidence scale was defined probabilistically, rather than qualitatively as in the latest uncertainty guide; however an explicit interpretation of probability (i.e. a subjective interpretation) was also provided, again in stark contrast to the AR5 uncertainty guide, which does not explicitly define probability anywhere.

It is also interesting to notice both the resemblance and *discrepancy* between the table of 'supplemental qualitative uncertainty terms' provided by the AR3 guide (Figure 2.2) and the 'depiction of evidence and agreement statements and their relationship to confidence' that one finds in the AR5 uncertainty guide (Figure 1.1). They resemble each other insofar as both depict the evidence and agreement as distinct dimensions. A crucial difference, however, is that although in this uncertainty guide the evaluation of the evidence dimension seemed to depend on just the amount of evidence underlying a particular finding, in the AR5 uncertainty guide the evidence dimension further depends on *additional* criteria,

such as consistency and independence of the available evidence. This *discrepant* resemblance may shed some light on the ultimate reason for the puzzling bifurcation of evidence and agreement in the evaluation of confidence that one finds in the AR5 uncertainty guide (discussed in Section 1.3): mere carelessness. That is, in light of this discrepant resemblance, it is hard to resist making the following conjecture: over the course of the IPCC assessment reports, revisions were made to the assessment criteria for the evaluation of evidence as it became clear that considerations on the 'amount of evidence' underlying a finding were rather unhelpful if not taken together with other considerations such as, for instance, independence and consistency of the various different lines of evidence underpinning that finding. However, although the introduction of these additional criteria for the evaluation of the evidence dimension makes it rather unclear how the agreement dimension can feasibly be independent from it, someone simply forgot to worry about this. In other words, the problematic bifurcation between evidence and agreement in the AR5 uncertainty guide may be simply due to an ill-thought-out stopgap bifurcation that first appeared in the AR3 uncertainty guide, and that got petrified in the process despite its irreconcilability with the revisions to the IPCC uncertainty guides later introduced.

Moss and Schneider's uncertainty guide was a first, admirable, attempt to provide the IPCC Working Groups with a single uncertainty framework for the communication of uncertainty in their findings with the objective of encouraging a more transparent and consistent treatment of uncertainty in the IPCC reports. Given this, one may wonder what actually happened in the AR3. To what extent was this guide successful in achieving its objective?

As it turns out, despite the presentation of this calibrated uncertainty language, in the AR3 itself there was nonetheless quite a large discrepancy between WG I and II in their definition and use of standard terms to describe uncertainty.[1] As Manning (2006) notes, an analysis of the language used by the two Working Groups readily reveals this discrepancy: whereas WG II followed the directions of the guidance in its usage of the confidence metric to represent the authors' degree of confidence in key findings, WG I actually introduced a *new*

[1] WG III did not use calibrated uncertainty language in its contribution to the AR3.

uncertainty metric in the AR3 and used this metric to characterise uncertainty in their findings: it was called 'likelihood'.[2]

According to Manning (2006), the reason behind this addition was not obscure. Many of the key findings found in the literature consulted by WG I were supported by large collections of data and hence the authors of WG I often felt they could simply rely on statistical analysis as opposed to expert judgment to estimate the probability of the occurrence of an event. As mentioned previously, Moss and Schneider (2000) explicitly recommended a subjective interpretation of probability; curiously, though, they also seemed to have anticipated that in some cases (e.g. when dealing with 'a long sequence of observational records, replicable trials, or model runs') authors might choose to adopt a 'frequentist' approach to characterize uncertainty in their findings, as opposed to relying on expert judgment. And while they did suggest that it is not always feasible to adopt a frequentist approach to characterize uncertainty, they did not in fact discourage authors from using such a frequentist approach if deemed appropriate, as long as the approach used was made explicit:

> [*A*]*uthors should explicitly state what sort of approach they are using in a particular case*: if frequentist statistics are used the authors should explicitly note that, and likewise if the probabilities assigned are subjective, that too should be explicitly indicated. Transparency is the key in all cases. (Moss and Schneider 2000, 36; emphasis in the original)

Arguably, the absence of further guidance as to how to explicitly indicate the chosen approach, led the WG I authors to take it upon themselves to decide how to do so: namely, by introducing an additional likelihood scale. In any case, during the IPCC workshops in preparation for the Fourth Assessment Report (AR4) the discrepancy between WG I and WG II's usage of calibrated uncertainty terms did not go unnoticed. Below is a passage from the resulting concept paper:

[2] In the summary for policy makers, WG I defined the likelihood scale as follows: 'The following words have been used where appropriate to indicate judgmental estimates of confidence: *virtually certain* (greater than 99% chance that a result is true); *very likely* (90%-99% chance); likely (66%-90% chance); *medium* likelihood (33%-66% chance); *unlikely* (10%-33% chance); *very unlikely* (1%-10% chance); *exceptionally unlikely* (less than 1% chance)'. (IPCC 2001, 2)

> The different approaches taken in the [AR3] by WGs I and II highlights some implications of choice of language. The WG I use of likelihood as a basis for approaching uncertainty focuses on probability of outcomes, and was clearly intended to be interpreted that way despite the definition in the WG I Summary for Policymakers referring to 'judgmental estimates of confidence'. The WG II use of level of confidence focused on degree of understanding and consensus, but at times was used as a proxy for the probability of an outcome. *In retrospect both likelihood and level of confidence may need to be addressed and the language used should not confuse the two*. (Manning et al. 2004, 6; my emphasis)

Although the nature of the ideas expressed in this passage is, in my view, far from clear (I will discuss why this is in Section 2.4), what *is* clear is that WG I's decision to introduce the likelihood scale in the AR3 was neither unnoticed nor rebuked. Indeed, as we will see in the next section, it was decided that the likelihood scale was there to stay . . .

2.3 The second IPCC uncertainty guide (for the AR4)

The uncertainty guide for the AR4 (IPCC, 2005) provided not one, but two quantitative scales, one for confidence and one for likelihood. The confidence scale, again, had five levels ranging from "very low confidence" to "very high confidence". These levels were now defined in terms of 'chance of being correct' as shown in Figure 2.3.

Terminology	Degree of confidence in being correct
Very High confidence	At least 9 out of 10 chance of being correct
High confidence	About 8 out of 10 chance
Medium confidence	About 5 out of 10 chance
Low confidence	About 2 out of 10 chance
Very low confidence	Less than 1 out of 10 chance

FIGURE 2.3: 'Quantitatively calibrated levels of confidence' (IPCC 2005, 3)

The authors were advised to use this confidence scale 'to characterize uncertainty that is based on expert judgment as to the correctness of a model, an analysis or a statement' (ibid., 3).

The likelihood scale had seven levels which ranged from "exceptionally unlikely" to "virtually certain". These were defined in terms of 'probability of occurrence' as shown in Figure 2.4.

Terminology	Likelihood of the occurrence/ outcome
Virtually certain	> 99% probability of occurrence
Very likely	> 90% probability
Likely	> 66% probability
About as likely as not	33 to 66% probability
Unlikely	< 33% probability
Very unlikely	< 10% probability
Exceptionally unlikely	< 1% probability

FIGURE 2.4: 'Likelihood Scale' (ibid., 4)

The authors were advised to use this likelihood scale to express 'a probabilistic assessment of some well-defined outcome having occurred or occurring in the future', which 'may be based on quantitative analysis or an elicitation of expert views' (ibid., 4).

Finally, similarly to the AR3 guide, qualitative language was also presented for the authors to characterize the amount of evidence and the level of agreement/consensus underlying a finding. With regards to this language there were, however, two differences from the previous uncertainty guide. First, while the AR3 guide provided four qualitative terms to together characterize the amount of evidence and the level of agreement/consensus, the AR4 guide now advised the authors to describe evidence and agreement/consensus separately, with now nine possible ways to combine the summary terms for evidence and agreement, as shown in Figure 2.5. This was perhaps due to the AR3's four qualitative terms (*Speculative*, *Competing Explanations*,'*Established but Incomplete* and *Well-established*) not having been found sufficiently nuanced to adequately characterize the amount of evidence and level of agreement or consensus.

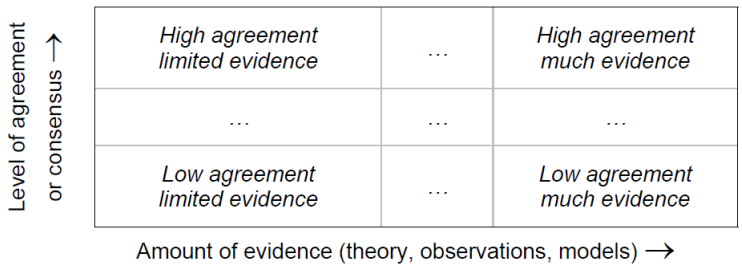

FIGURE 2.5: 'Qualitatively defined levels of understanding' (ibid., 3)

The second, more striking difference, however, was that this qualitative language was now no longer presented as a *supplementary* language to be used *together* with confidence to explain the reasoning behind a particular level of confidence, as in the previous AR3 guide. Rather, the authors were now instructed to use these qualitative terms

> to summarize judgments of the scientific understanding relevant to an issue, or to express uncertainty in a finding *where there is no basis for making more quantitative statements*. (ibid., 3, my emphasis)

In other words, these qualitative terms were now treated as *replacements* for confidence and likelihood terms, contrary to what was recommended in the AR3 uncertainty guide; only in cases of "high agreement much evidence" were the authors encouraged to characterize uncertainty using the confidence and likelihood scale provided (ibid., 3).

Given this substantially different uncertainty framework, one may wonder what happened in the AR4 report this time. How did the AR4 authors interpret and use the confidence and likelihood metrics provided by this guide? Mastrandrea and Mach's (2011) analysis of the practice of the AR4 authors reveals things got *really quite messy*. Sometimes confidence and likelihood terms were used together in a statement, consistent with an interpretation of likelihood and confidence as representing different aspects of uncertainties (whatever those may be). Other times only likelihood terms were used in a statement, consistent with an interpretation of the likelihood metric as encompassing all relevant uncertainties (though in practice, it was not always clear its usage did achieve as much). Other

times still, the likelihood or confidence metrics were used interchangeably, thus obliterating any conceptual distinction between them.

Arguably, this mess should have been expected: several aspects of this guide left the understanding of what the confidence and likelihood scales actually represented rather open to the AR4 authors' interpretation. For a start, recall that confidence levels were now defined in terms of chances. But if the confidence scale was meant to be used 'to characterize uncertainty that is based on expert judgment as to the correctness of a model, an analysis or a statement', then it is hard to see what role chance would have to play in the characterization of confidence. For instance, it is hard to see what it would mean for a model or an analysis to have at least 9 out of 10 chances of being correct. The use of the word 'chance' in the definition of confidence, without any further qualification, could have been (and arguably was) a source of misinterpretation as to what the confidence metric was meant to characterize. Recall also that that the summary terms for evidence and agreement/consensus were now presented as *replacements* for confidence, rather than supplementary. But then the relationship between the evaluation of evidence and agreement/consensus on the one hand and that of confidence on the other was somewhat unclear; in particular, it seems very hard to reconcile the above interpretation of confidence with the AR4 guide's directions to use confidence only in cases where there was 'high agreement much evidence'. Surely the experts' judgment about the correctness of a statement must have depended on the level of agreement and the amount of evidence available. But if the authors were encouraged to report confidence only in cases with 'high agreement much evidence', then it is hard to see how, under this interpretation of the confidence metric, the IPCC authors would ever be able to use the confidence terminology not at the top of scale; that is, how they could ever talk about 'very low confidence' or 'low confidence' or even 'medium confidence'.

This inconsistent treatment and communication of uncertainty across the AR4 did not go unnoticed; crucially, the acknowledgment that this inconsistency was largely due to a lack of clarification about the distinction between the confidence and the likelihood metrics seems to have been an important motivation for producing a new uncertainty guide for the AR5 (IPCC, 2010). Here is a

passage from Annex A of the AR5 guide:

> Consistent treatment and communication of uncertainty across the Working Groups is a key cross-cutting issue for the IPCC and goal for the AR5. To address this important issue, the Co-Chairs of the three Working Groups convened a small meeting 6-7 July 2010 at the Jasper Ridge Biological Preserve in Stanford, CA, USA. The outcome of the meeting was a decision to produce updated Guidance Notes for AR5, *with the goal of improving the distinction and transition between different metrics and their consistent application across the Working Groups in the AR5*. (IPCC 2010, Annex A; my emphasis)

It is evident that in the hope of clarifying the distinction between the confidence and likelihood metrics in the AR5, an attempt was made to deal with some of the puzzling aspects involved in the characterization of confidence that I mentioned above. Indeed, as seen in Section 1.2, the AR5 uncertainty guide no longer defines confidence in terms chances. Rather confidence is presented as a qualitative scale whose evaluation now explicitly depends on the evaluation of 'evidence' and 'agreement' as Figure 1.1 illustrates (although of course, as argued in Section 1.3, the nature of this dependency is still far from clear).

However, I believe there was a much more fundamental problem that came with the emergence of an additional likelihood metric which failed to be fully appreciated at the time, and that can provide an important insight into why, despite the efforts put into improving the characterization of confidence in the AR5 uncertainty framework, the relationship between confidence and likelihood is still not quite as unambiguous as it should be. I do not believe this problem lies in the emergence of the likelihood metric per se, but rather in the reason(s) for its emergence. That is, the problem lies in the fact that there were two, related but distinct, reasons for its emergence that got somehow merged into one. Why do I think this is a problem? Because as these two reasons came to merge and blur one into the other, the concept of probability came to be treated without the rigour and care it necessitates, leaving its interpretation impossible – for the IPCC authors, for me, or for whoever else cared/cares to try.

2.4 On the reason(s) for the emergence of two uncertainty scales

Probability is undeniably a very useful concept for the characterization of uncertainty, but it is also a meaningless one when used without a clear and understandable interpretation of it. As we will see, in this section, the IPCC has made various remarks concerning probability, but (alas) lack of sufficient rigour and precision makes those remarks very hard to interpret and understand. For the purpose of disambiguating some of those remarks, I will have to make three assumptions about the interpretations of probability, assumptions that I think the IPCC has also made, if not always explicitly:[3]

1. I will assume that there are two kinds of probability: subjective probabilities and objective probabilities: whereas subjective probabilities (also known as credences or degrees of belief) depend on the mental states of individual agents, objective probabilities (also known as chances) are features of the mind-independent world;

2. I will assume that it is possible to make true chance statements about a system that obeys deterministic laws, that is I will assume that chances and determinism are compatible;

3. I will assume that the IPCC has a conceptually coherent set of methods for producing objective probabilities (or an estimate of them).

As discussed in Section 2.2, during the IPCC workshops in preparation for the Fourth Assessment Report (AR4), WG I's decision to introduce and use a new uncertainty scale (i.e. the likelihood scale) in the AR3 in addition to the one presented in the AR3 uncertainty guide (i.e. the confidence scale) did not go unnoticed. And, as we have seen Section 2.3, it was in the end decided that the new uncertainty guide (for the AR4) would have to include *both* a confidence and a likelihood metric. But what was the reason behind this decision? I believe there were in fact two distinct reasons.

[3]The first two assumptions are fairly uncontroversial; the last one in my view less so, as I will discuss in Section 2.5.

To understand the first reason, consider again the following passage from the AR4 workshop concept paper:

> The different approaches taken in the TAR by WGs I and II highlights some implications of choice of language. The WG I use of likelihood as a basis for approaching uncertainty focuses on probability of outcomes, and was clearly intended to be interpreted that way despite the definition in the WG I Summary for Policymakers referring to 'judgmental estimates of confidence'. The WG II use of level of confidence focused on degree of understanding and consensus, but at times was used as a proxy for the probability of an outcome. In retrospect both likelihood and level of confidence may need to be addressed and the language used should not confuse the two. (Manning et al. 2004, 6)

This passage acknowledges that WG I and WG II took distinct approaches in the characterization of uncertainty and suggests that both a confidence and a likelihood metric might therefore be needed to distinguish between them. Indeed, as mentioned above, whereas WG I heavily relied on a frequentist approach in the characterization of uncertainty in the AR3, WG II largely relied on a subjective approach instead. So although probability is not defined, I think the most plausible interpretation of this passage is what I will call **reason 1**: there is a distinction between subjective and frequentist/objective probabilities, and hence we need two distinct uncertainty scales, confidence and likelihood, to distinguish between them.

Consider now the following two passages both reflecting on why one uncertainty scale was deemed insufficient for the IPCC authors to adequately characterize uncertainty in their findings. The first is from Manning himself, whereas the second passage appears in the AR4:

> The wide ranging and inter-disciplinary discussion of uncertainty and risk that took place during preparations for the AR4 has led to a richer language and more comprehensive structure for determining and describing uncertainties. While these approaches are clearly

> rooted in the Guidance Paper for the TAR,[4] *they also reflect a real evolution in our thinking and one of their key results has been to draw out more clearly the distinction between the assessed likelihood of specific outcomes and the confidence that the science community has in its ability to determine such likelihood*. (Manning 2006, my emphasis)

> The uncertainty guidance provided for the [AR4] draws, for the first time, a careful distinction between levels of confidence in scientific understanding and the likelihoods of specific results. *This allows authors to express high confidence that an event is extremely unlikely (e.g., rolling a dice twice and getting a six both times), as well as high confidence that an event is about as likely as not (e.g., a tossed coin coming up heads)*. Confidence and likelihood as used here are distinct concepts but are often linked in practice. (AR4, 22, my emphasis)

These passages seem to articulate a rather different idea from that expressed in the previous one, which I'll call **reason 2**: An adequate communication of uncertainty necessitates two uncertainty scales: a likelihood scale to indicate the assessed probability of an event, and a confidence scale to indicate the IPCC authors' confidence in their ability to determine it.

Reason 1 and **reason 2** are not the same reason, and yet they both seem to have played a role in the emergence of two uncertainty scales in the AR4 uncertainty framework. Let us try to better understand what I have called **reason 1**, the need to distinguish between subjective and frequentist/objective probabilities. Although having two distinct uncertainty scales would seem to address this need, one may well ask: why the need to have two different uncertainty scales in order to distinguish these two kind of probabilities? Of course, as the first uncertainty guide itself remarked 'transparency is key' – but it does seem a rather big leap to go from agreeing with this, to the introduction of an additional likelihood scale altogether. Why take this leap? Consider the 'principal principle' (Lewis, 1980), according to which a rational agent should always conform their

[4] TAR is short for 'Third Assessment Report', to which I refer throughout the paper as AR3.

degrees of belief to the objective probabilities of the occurrence of an event.[5] In other words, if one knows the objective probability that an event will occur then one's degree of belief in the occurrence of that event should be the same. In light of this principle, if the probabilities derived using a frequentist approach are what the IPCC authors believe to be the objective probabilities of the occurrence of an event, it seems plausible that those probabilities should also be the same as the IPCC authors' degrees of belief in the occurrence of that event. But then, it is not clear why, under **reason 1**, two uncertainty scales would strictly be needed for an adequate communication of uncertainty.

This brings us to what I called **reason 2**: the need to distinguish the assessed probability of an outcome and the *confidence* that the science community has in its ability to determine it. The recognition that the objectivity of those assessed frequentist probabilities is always conditional on various assumptions together with the possibility that the science community might not have much confidence in these assumptions, seems to be the very reason why those probabilities cannot be interpreted as the degrees of belief of the science community. But if those assessed frequentist probabilities might not be *objective* probabilities after all, then what *are* they? How should one *interpret* those probabilities? And most crucially why rely on a frequentist approach to quantify uncertainty in the first place if the confidence in the assumptions that justify the use of a frequentist approach are not sufficiently high? Without an unambiguous answer to these questions, things are bound to become very confusing. Why? Because it is no longer clear how the concept of probability is used and defined, leaving its interpretation impossible.

The current AR5 uncertainty guide's recommendation to only assign 'likelihood' in cases where 'confidence' is sufficiently high is, in my view, an attempt to deal with this tension. The underlying thought behind this recommendation is, arguably, the following: if the assessed probability is to be 'objective', then likelihood should only be used if there is sufficiently high confidence in the assumptions that justify the use of a frequentist approach. But this is evidently not

[5] As long as the agent does not have inadmissible knowledge about the occurrence of that event.

an adequate attempt to deal with this tension since, as argued in Section 1.4, this recommendation is *incompatible* with any possible interpretation of 'likelihood' and 'confidence'.

2.5 Likelihood revisited: objective probabilities, subjective probabilities or neither?

In the previous section, I have argued that one of the reasons behind the decision to include both a confidence and a likelihood metric in the IPCC uncertainty framework (**reason 1**) was to distinguish frequentist approaches from subjective approaches in the characterization of uncertainty, which I interpreted as the need to draw a distinction between subjective probabilities and objective probabilities. However, in this section, I will have a close look at some common frequentist approaches in the characterization of uncertainty by the IPCC and I will conclude that they are not conceptually coherent methods for producing objective probabilities. This, I will argue, raises doubts as to whether my interpretation of **reason 1** behind the decision to include both a confidence and a likelihood metric in the IPCC uncertainty framework is an accurate interpretation after all.

But before I get to all that, it will be helpful to give a quick review of some basic frequentist statistical notions.

2.5.1 A Review of Some Statistical Concepts

In this subsection, I will give a quick overview of some basic frequentist notions in statistical inference. In particular, I will discuss the well-known notion of a confidence interval and the less well-known notion of a tolerance interval.

Confidence intervals

A common aim in statistical inference is to estimate unknown parameters that characterize a population of interest; and relatedly, to assess the (unavoidable) uncertainty in those estimations. To understand how this objective is carried out in a frequentist paradigm, it is important first to understand the interpretation of

probability under it. According to the frequentist view of probability, an event's probability is the limit of its relative frequency as the number of trials increases. Under this view, the probability of an event is objective since it can, in principle, be found by a repeatable objective process. This interpretation of probability is in stark contrast with a Bayesian view of probability, according to which an event's probability represents the degrees of belief of an agent in that event.

Before I can explain how the frequentist view of probability relates to the assessment of uncertainty in the estimation of a parameter of a population, I must first introduce a few key statistical concepts. In statistics, one usually takes a sample from a population of interest in order to estimate the properties of that population. A sample statistic is a mathematical function of the sample, whereas a parameter is any numerical quantity that characterizes some aspect of the population of interest. In other words, the value of sample statistics are the things one can calculate from one's sample, and the value of the population parameters are the things that one is trying to learn about. Evidently, a sample statistic and a population parameter are conceptually distinct concepts, but what links them together is the following: sample statistics can be used to *estimate* population parameters.

How can a sample statistic be used to estimate a population parameter? The notion of a sampling distribution is key here. As already mentioned, a sample statistic is a function from the sample and hence its value can be calculated from the sample we have taken. But in principle, there is nothing stopping us from taking another sample of the population and calculating the value of the sample statistic one more time, which could be different from the one obtained earlier. And, still in principle, one could do this yet another time and so on. The sampling distribution is the probability distribution of a sample statistic obtained from taking an infinite number of samples of the same size from a population. Of course, statisticians can't take an infinite number of samples, and as a matter of fact they usually just take one. However, the fact that they can't, is in the frequentist paradigm, no reason to reject the objectivity of the sampling distribution: the sampling distribution is real and although one might not be able to find it in practice, one can often find it theoretically.

How does one find sampling distributions theoretically? It is not always easy, but mathematics often comes to the rescue. A very important mathematical theorem, for instance, is the *Central Limit Theorem*, in light of which one can assume that the sum of a large number of independent random variables, each with finite mean and variance, will be approximately normally distributed, irrespective of the distribution function of the random variables. In particular, the Central Limit Theorem tells us that if a population has mean μ and standard deviation σ and one takes sufficiently large samples from that population of size N, the sampling distribution of the sample mean is approximately normally distributed with mean μ and standard deviation $\sigma/\sqrt{N}$ (also known as the standard error). What is very appealing about the Central Limit Theorem is that it holds regardless of whether or not the population of interest is normally distributed. Hence, because of the Central Limit Theorem, statisticians can often assume that the sampling distribution of the mean is normal, despite not knowing the actual shape of the population distribution.

But why should one care about the sampling distribution? One reason one should care (under a frequentist paradigm) is that if one knows the shape of the sampling distribution, one can use this information to obtain a *confidence interval*. A confidence interval is an interval estimate of a population parameter with an associated confidence level, where the confidence level represents the limit of the relative frequency of the confidence intervals that will contain the true value of the unknown population parameter. Below is a simple example that illustrates how knowing the sampling distribution of a statistic allows one to find the confidence interval for a parameter of interest.

Suppose we want to estimate the mean height of all British 8 year old girls. So the population of interest consists of all British 8 year old girls and we want to learn the mean height μ of this population. Assume further that we happen to already know the value of standard deviation of the population σ. We decide to take a large random sample of size N and calculate the mean height of this sample $X = x$. By appealing to the Central Limit Theorem, we are happy to assume in this case that the sampling distribution of the sample mean X is normally distributed with mean μ and standard error $\frac{\sigma}{\sqrt{N}}$. From this we can

conclude that the interval $\mu \pm 1.64\frac{\sigma}{\sqrt{N}}$ includes 90% of the sampling means X's in repeated sampling as shown in the figure below.[6]

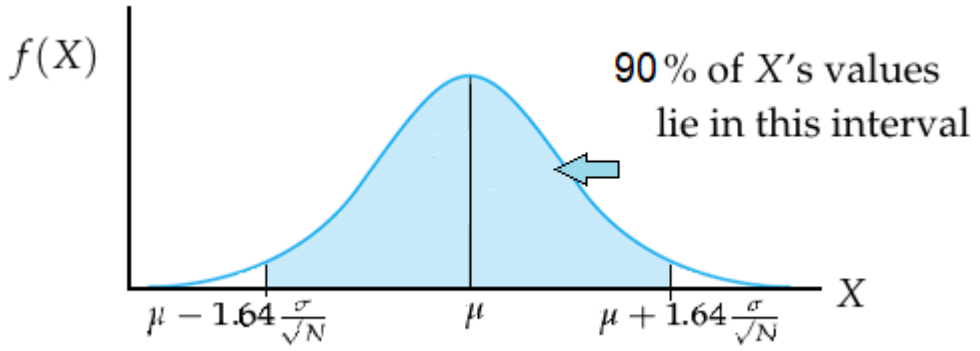

FIGURE 2.6

Notice further that whenever the value of the sample mean X falls in the interval $\mu \pm 1.64\frac{\sigma}{\sqrt{N}}$, the interval $X \pm 1.64\frac{\sigma}{\sqrt{N}}$ will contain the parameter μ as shown in the figure below.

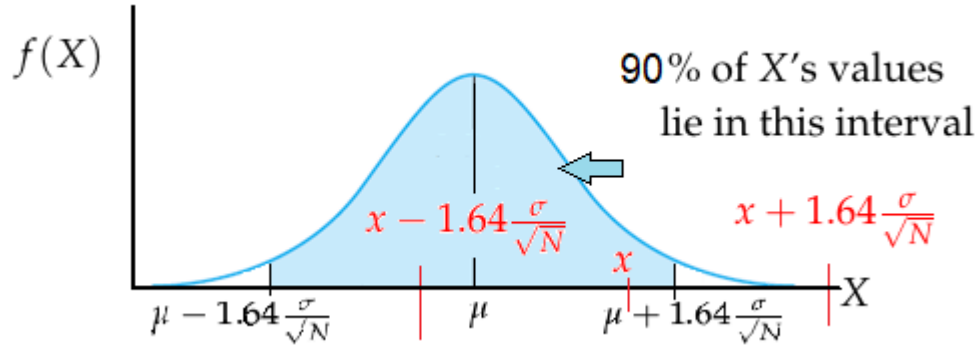

FIGURE 2.7

But this means that if confidence intervals $X \pm 1.64\frac{\sigma}{\sqrt{n}}$ are constructed from an infinite number of independent sample statistics, 90% of those intervals will contain the true value of the parameter μ. Hence the interval $x \pm 1.64\frac{\sigma}{\sqrt{n}}$ is the 90% confidence interval for the mean height μ of all British 8 year old girls.

A few words on the interpretation of a confidence interval are in order. A specified level of confidence does not refer to the confidence interval that has been computed, but rather it refers to the *procedure* which has been used to construct that confidence interval. So a 90% confidence interval for an unknown

[6] 90% of the area under a normal curve lies within roughly 1.64 standard deviations of the mean.

parameter of a population (e.g. the mean height μ in the example above) can be interpreted as follows: "If one repeatedly calculates 90% confidence intervals from independent random samples of the same size, 90% of these intervals would, in the long run, correctly include the actual value for μ. Hence from the frequentist definition of probability, it follows that: there is a 0.9 probability that a 90% confidence interval calculated from a random sample will contain μ. However, and crucially, this does *not* entail that the probability of a *specific* 90% confidence interval contains μ is also 0.9. In a frequentist paradigm, once the confidence interval is calculated, it either includes the correct parameter value or it doesn't. Hence whether or not that confidence interval contains the true parameter value is not a matter of probability.

In light of this, one might ask: why should we care about the notion of a confidence interval if as soon as we calculate one, all we can say about it is that it either does or doesn't cover the correct parameter value of the population in question? According to a 'coverage probability rationale' (Mayo 2018, 193), one would argue that we should trust a confidence interval to include the correct parameter value simply because, for instance, a 90% confidence interval will correctly cover the true parameter value 90% of the time in repeated use. Hence we should rely on confidence intervals to estimate a population's parameter because they are generated from a procedure that performs well when used repeatedly. However, as Mayo argues this performance oriented justification for relying on confidence intervals is unsatisfactory for it is not clear how knowing the performance of a procedure in repeated use can help us evaluate how it is performing *now*. For instance, one may be tempted to say something like: 'if a procedure is rarely wrong we may assign a low probability to its being wrong in the case at hand'. But this would be 'dangerously equivocal, since the probability properly attaches to the method of inferring' (Mayo 2018, 194). According to Mayo, the ultimate justification for relying on a given confidence interval to estimate a parameter value is that the hypothesis that the parameter value lies in the calculated confidence interval has passed a severe test. That is, the justification for relying on confidence intervals is counterfactual. Since this is a two-tail confidence interval there are two counterfactual claims we care about:

1) Were μ lower than the lower confidence limit then it is very probable (a probability greater than 0.95) that the procedure would have yielded a smaller sample mean than the one observed and 2) Were μ greater than the upper confidence limit then it is very probable that the procedure would have yielded a larger sample mean than the one observed.[7]

Leaving aside questions concerning the interpretation of confidence intervals and why we should rely on them for estimating parameters of a population, there are few technical remarks that are worth making. In the example above I assumed that we already know the standard deviation of the population. This assumption allowed me to compute a confidence interval by relying on the Central Limit Theorem. However, in most realistic cases we do not already know the standard deviation of a population, hence we can't rely on the Central Limit Theorem to infer the sampling distribution of the sample mean. Furthermore, the Central Limit Theorem only holds for sample sizes that are sufficiently large. Hence, whenever the sample size is not considered to be sufficiently large, we also can't rely on the Central Limit Theorem to infer the sampling distribution of a sample statistic. In these cases we'll have to rely on some other mathematical theorem. For instance, it can be shown that if the population of interest is normally distributed, the t-statistic of a random sample of size N:

$$\frac{X - \mu}{S/\sqrt{N}}, \tag{2.1}$$

where X is the mean of the sample and S is the standard deviation of the sample, has a student's t-distribution with N-1 degrees of freedom.[8] Hence in many cases, if a sample is not sufficiently large or if the standard deviation of the population is unknown, we can nonetheless use this mathematical fact to derive a confidence interval for μ.

[7] These two counterfactual claims follow from the duality between confidence interval estimation and tests (see Mayo 2018, 190-93). For instance, the hypothesis that μ is less than the lower limit of a calculated 90% confidence interval is rejected at a p-value of 0.05. I will come back to Mayo's notion of Severity in Chapter 6.

[8] The shape of student's t-distribution resembles the bell shape of the normal distribution with mean 0 and variance 1, but has fatter tails. However, as the number of degrees of freedom increases, the t-distribution approaches the normal distribution with mean 0 and variance 1. See, for instance Mason et al. (2003, 46-47) for a quick introduction to the student's t-distribution and when it can be reasonably assumed to be an adequate approximation to the normal distribution.

Tolerance intervals

As discussed above, a confidence interval of a parameter represents an estimate of *that* parameter. For instance, the 90% confidence interval for the mean μ in the example above, is an estimate of μ and μ alone. Hence the size of a confidence interval is exclusively due to the sampling error and will approach a zero-width as the sample size increases. Hence, a confidence interval for say a population's mean height says nothing about how that population deviates from that mean. But suppose I am interested in estimating the interval of height values that includes e.g. 90% of the population's heights. In this case, I don't want a confidence interval; what I want is a *tolerance interval*.

A *tolerance interval* is an interval that contains a specified proportion of a population, at a specified level of confidence.[9] Evidently, when it comes to calculating a tolerance interval the most straightforward cases are those in which the parameters of the population of interest's distribution are known. In those cases, it is straightforward to compute the interval that includes a specified proportion of the population's distribution. For instance, let's return to the example above. Suppose we know that the distribution of the population of British 8 year old girls is normal with mean μ and standard deviation σ. In this case, one can easily construct a tolerance interval to include, say 90% of the population: it will be the interval $\mu \pm 1.64\sigma$ (i.e. it will be the interval included within the 5% and 95% percentile of the population). No "confidence" is attached to this interval, because all the parameters of the population's distribution are known. Hence, in light of the frequentist interpretation of probability, one could make the following statement in this case: there is a probability of 0.9 that the height of a randomly chosen British 8 year old girl falls in the interval $\mu \pm 1.64\sigma$ (since the limit of the relative frequency of the event ' a randomly chosen British 8 year old girl falls in the interval $\mu \pm 1.64\sigma$' as the number of trials increases is equal to 0.9).

However, in most realistic cases the parameters of the population distribution are not known (e.g. all cases where the available information is limited

[9] Where again, the specified level of confidence refers to the procedure used for constructing the tolerance interval, not to the tolerance interval that has been computed.

to the sample we have taken). Hence most applications of tolerance intervals require the estimation of population's mean and standard deviation. In these cases, we will have to construct a tolerance interval at the chosen confidence level.[10] So for instance a (95%, 90%) tolerance interval for the height of a British 8 year old girl tells us, at a 95% confidence level, that the height of at least 90% of British 8 year old girls falls within that interval. Hence, the width of the tolerance interval is affected by both the population's proportion we want to cover and the desired level of confidence. This means that, whereas a confidence interval's size is entirely due to sampling error, and will approach a zero-width interval at the true population parameter as sample size increases, a tolerance interval's size is due partly to sampling error and partly to the actual variance in the population, and will approach the population's probability interval as the sample size increases.

Calculating tolerance interval can be rather tricky, and since this is not a PhD on statistical methods, I will spare the reader with the details! However, I'd like to give a very quick practical example to show how the choice of confidence can greatly affect the width of the tolerance interval that one will obtain.[11] Suppose we have a sample of 30 randomly chosen British 8 year old girls. The mean height of the sample is 130 cm and the standard deviation is 4 cm. Suppose further that we know that the height distribution of the population of all the British 8 year old girls is normal. Below are 3 distinct tolerance intervals covering the same proportion of the population (i.e. 90%) but at distinct levels of confidence:

- A (90%, 90%) tolerance interval in this case would be the interval [121.9 cm, 138.1 cm];
- A (95%, 90%) tolerance interval in this case would be the interval [121.4 cm, 138.6 cm];
- A (99%, 90%) tolerance interval in this case would be the interval [120.4 cm, 139.5 cm].

[10] See, for instance, Mason et al. (2003, 50-52) for a more detailed introduction to tolerance intervals.

[11] I have relied on a statistical interactive page very kindly provided by Pezzullo (2005) to calculate the tolerance intervals below.

As this example illustrates, as we increase the level of confidence with which we want to cover 90% of the population's heights, the width of the tolerance interval increases substantially. Finally, it is worth mentioning that had we simply assumed that the sample mean (130 cm) and standard deviation (4 cm) are sufficiently accurate estimates for the population mean and standard deviation and derived a probability distribution for the population height from this assumption, we would have concluded that the interval covering 90% of the population's heights is the interval $[130 \mp 1.64 \times 4 \text{ cm}] = [123.4 \text{ cm}, 136.6 \text{ cm}]$ (with no confidence attached), which is substantially smaller than any of the three tolerance derived above. Hence, and especially in cases where our sample is not very large, it is not reasonable to assume that the sample mean and standard deviation are sufficiently accurate approximations for the mean and standard deviation of the population of interest and derive a tolerance interval based on this assumption.

Populations

Finally, a few remarks on the very concept of a population are in order. As discussed above, in statistics one often takes a random sample from a population with the objective of learning something about that population. In many cases, as in the example above, the population consists of physical objects (e.g 8 year old girls living in the UK) with a particular characteristic (e.g. height) that we want to learn about. But there are many cases in which the population whose characteristics we are trying to learn about does not consist of physical objects. In the problem of repeated *measurement* of a quantity, for instance, the measurements are regarded as a sample from the population that would exist if the repetition could be continued indefinitely.

Suppose, for instance, that we are trying to measure the height of a tall building h and we expect the measurements (e.g. the readings on a very long meter stick) to approximate the quantity while not being exactly equal to it, due to some errors in our measurements. In this case the population of interest (often referred to as the parent population) consists of the infinite set of

measurements that could be taken of which our finite sample of measurements $(x_1, x_2, x_3, \ldots, x_n)$ is but a subset. In this case, we might be interested in finding the mean μ of this population insofar as we think that the mean of the parent population has the same value as the height of the building h (i.e. $\mu = h$).[12] And if the sample size is sufficiently large and we know the standard deviation of the parent population, then we can appeal to the Central Limit Theorem in this case and happily assume the sampling distribution of the mean is normal and derive a confidence interval for $\mu = h$. It is also worth noting that in the problem of repeated measurement, it is often assumed that the parent population distribution is itself normal.[13]

2.5.2 A closer look at likelihood: "multi-model ensemble methods" and a questionable desire for "objectivity".

Consider the following finding from the IPCC summary for policy makers:

> Increase of global mean surface temperatures for 2081–2100 relative to 1986–2005 is projected to *likely* be in the ranges derived from the concentration-driven CMIP5 model simulations, that is, 0.3°C to 1.7°C (RCP2.6) [. . .] (*very high confidence*) (SPM, 20)

In this section, I will discuss the methods used by the IPCC authors to characterize uncertainty in this finding (hereinafter referred to as "multi-model ensemble methods"). I will argue that they are not conceptually coherent methods for producing objective probabilities.[14] The multi-model ensemble methods I will discuss in this section are not the only methods the IPCC authors use to produce probabilities in their findings, but they are nonetheless fairly common; hence I

[12] But note that if there are systematic errors, that is errors that systematically cause the measured quantity to be shifted away from the real height of the building h (e.g. measurements with a cold meter ruler which appear bigger because the scale has contracted), it would be unreasonable to assume that the mean of the parent population has the same value of the quantity h.

[13] Again by appealing to the Central Limit Theorem: if a measurement result is simultaneously influenced by many uncertainty sources then if the number of the uncertainty sources approaches infinity the distribution function of the measurement result approaches the normal distribution, irrespective of the distribution functions of the factors/parameters describing the uncertainty sources.

[14] Although Winsberg (2018) has recently argued for a similar conclusion, in this section I hope to show more forcefully why this is the case

think an assessment of those methods can provide us with some general lessons about the practice of the IPCC authors in their treatment of uncertainties.

To understand the methods used by the IPCC authors to arrive at the range [0.3°C, 1.7°C] for the increase of global mean surface temperatures for 2081–2100 relative to 1986–2005 (from now on I will denote this by GMST81-100) and the likelihood level (*likely*) assigned to it is crucial first to recognize that as far as climate modelling is concerned,

> [...] even for a particular question and set of processes, different models exist. Strictly they are incompatible; they cannot be true at the same time. But they are usually seen as complementary, because they represent different plausible (although not necessarily equally plausible) approximations to the target system, given some computational constraints, limited and uncertain observations, and incomplete understanding of all processes [...] For example, there are several ways to parameterize atmospheric convection, and no scheme is clearly superior to the others for all climatic states, and parameters are not well constrained. (Knutti 2018, 330)

Indeed, there is a great deal of uncertainty about how to adequately represent the climate system. Due to this uncertainty, it is often impossible to choose which model, out of the available ones, future climate change projections should rely. Hence, current projections of future climate change very often rely on more than a single model. The most recent Coupled Model Intercomparison Projection Phase 5 (CMIP5), for instance, was a huge collaborative effort, involving more than 20 climate modeling groups from around the world (Taylor et al. 2012, 486), to promote a standard set of model simulations whose outputs were then analysed by the AR5 authors to produce many of their findings.[15]

The fact that there exist several models relying on different plausible but incompatible assumptions about the climate system, is not in itself a fact that can be changed; hence, arguably, as Parker remarks,

[15]The Coupled Model Intercomparison Project is now in its 6th phase.

> Despite the fact that one must be careful when interpreting the results produced by multi-model ensembles, when it comes to addressing the global warming issue, the ensemble approach seems clearly better than the two most obvious alternatives, that is, relying on a single model and/or making no use of climate models until a single 'best' one can be identified. (Parker 2006, 361)

So far so good. The question I am concerned with, however, is the following: how do the AR5 authors interpret the results produced by a multi-model ensemble to arrive at the finding above? More specifically, how are the outputs of the CMIP5 multi-model ensemble used to arrive at the interval [0.3°C, 1.7°C] and the *likely* assignment to that interval?

Consider the following table provided in the AR5 (IPCC 2013, 1055).

Table 12.2 | CMIP5 annual mean surface air temperature anomalies (°C) from the 1986–2005 reference period for selected time periods, regions and RCPs. The multi-model mean ±1 standard deviation ranges across the individual models are listed and the 5 to 95% ranges from the models' distribution (based on a Gaussian assumption and obtained by multiplying the CMIP5 ensemble standard deviation by 1.64) are given in brackets. Only one ensemble member is used from each model and the number of models differs for each RCP (see Figure 12.5) and becomes significantly smaller after 2100. No ranges are given for the RCP6.0 projections beyond 2100 as only two models are available. Using Hadley Centre/Climate Research Unit gridded surface temperature data set 4 (HadCRUT4) and its uncertainty estimate (5 to 95% confidence interval), the observed warming to the 1986–2005 reference period (see Section 2.4.3) is 0.61°C ± 0.06°C (1850–1900), 0.30°C ± 0.03°C (1961–1990), 0.11°C ± 0.02°C (1980–1999). Decadal values are provided in Table AII.7.5, but note that percentiles of the CMIP5 distributions cannot directly be interpreted in terms of calibrated language.

	RCP2.6 (ΔT in °C)	RCP4.5 (ΔT in °C)	RCP6.0 (ΔT in °C)	RCP8.5 (ΔT in °C)
Global: 2046–2065	1.0 ± 0.3 (0.4, 1.6)	1.4 ± 0.3 (0.9, 2.0)	1.3 ± 0.3 (0.8, 1.8)	2.0 ± 0.4 (1.4, 2.6)
2081–2100	1.0 ± 0.4 (0.3, 1.7)	1.8 ± 0.5 (1.1, 2.6)	2.2 ± 0.5 (1.4, 3.1)	3.7 ± 0.7 (2.6, 4.8)
2181–2200	0.7 ± 0.4 (0.1, 1.3)	2.3 ± 0.5 (1.4, 3.1)	3.7 ± 0.7 (-,-)	6.5 ± 2.0 (3.3, 9.8)
2281–2300	0.6 ± 0.3 (0.0, 1.2)	2.5 ± 0.6 (1.5, 3.5)	4.2 ± 1.0 (-,-)	7.8 ± 2.9 (3.0, 12.6)
Land: 2081–2100	1.2 ± 0.6 (0.3, 2.2)	2.4 ± 0.6 (1.3, 3.4)	3.0 ± 0.7 (1.8, 4.1)	4.8 ± 0.9 (3.4, 6.2)
Ocean: 2081–2100	0.8 ± 0.4 (0.2, 1.4)	1.5 ± 0.4 (0.9, 2.2)	1.9 ± 0.4 (1.1, 2.6)	3.1 ± 0.6 (2.1, 4.0)
Tropics: 2081–2100	0.9 ± 0.3 (0.3, 1.4)	1.6 ± 0.4 (0.9, 2.3)	2.0 ± 0.4 (1.3, 2.7)	3.3 ± 0.6 (2.2, 4.4)
Polar: Arctic: 2081–2100	2.2 ± 1.7 (-0.5, 5.0)	4.2 ± 1.6 (1.6, 6.9)	5.2 ± 1.9 (2.1, 8.3)	8.3 ± 1.9 (5.2, 11.4)
Polar: Antarctic: 2081–2100	0.8 ± 0.6 (-0.2, 1.8)	1.5 ± 0.7 (0.3, 2.7)	1.7 ± 0.9 (0.2, 3.2)	3.1 ± 1.2 (1.1, 5.1)

FIGURE 2.8

In each slot of this table, the first entry shows the mean ± the standard deviation of the model ensemble's predictions for the selected period, region and RCP. The second entry (given in brackets) is an interval obtained by constructing a normal distribution with the same mean and standard deviation of that model ensemble, and by taking the 5-95% range (the mean ±1.64 standard deviation) of that distribution. Notice further that the interval circled in red $[0.3C, 1.7C]$ is the same interval that appears in the finding above. So what is going on here? As Winsberg succinctly explains,

> The most common method of estimating the degree of structural uncertainties in the predictions of climate models is a set of sampling methods called “multi-model ensemble methods". The core idea is to examine the degree of variation in the predictions of the existing set of climate models [. . .] By looking at the average prediction of the set of models and calculating their standard deviation, one can produce a probability distribution for every value that the models calculate. [. . .] If 80 percent of the results from a space of models [. . .] lie in the range, then the probability of the true result lying in that range is said to be 80 percent. (Winsberg 2018, 96)

And indeed, these “multi-model ensemble methods” described by Winsberg, are the very ones that are used by the IPCC authors to arrive the interval [0.3°C, 1.7°C] for the GMST81-100 in the above finding: the average (which is equal to 1°in this case) and the standard deviation (which is equal 0.4°in this case) of the available set of models' predictions for the GMST81-100 are used to produce a normal distribution for the GMST81-100. The interval $[0.3C, 1.7C]$ is then obtained from taking the 5-95% range of this probability distribution.

Of course, as the reader will have noticed, in this case the 5-95% range of this probability distribution is not assessed by the IPCC authors to be a *very likely* range for the GMST81-100, as it should be according to the calibrated uncertainty language for likelihood (see Figure 1.2), but as a merely *likely* range. This is explained in the summary for policy makers to be due to an ‘accounting for additional uncertainties or different levels of confidence in models’ (IPCC 2013, 23), a point to which I will come back at the end of this section. However, despite this subsequent downgrading of the likelihood assignment from *very likely* to *likely*, the assessed range is the very same as the one derived using the “multi-model ensemble methods" described by Winsberg. Hence these methods nonetheless play a crucial role in the characterization of this finding.

But what to think of these “multi model ensemble methods”? Why are the authors assuming (at least in the first instance in this case, before the downgrading of likelihood) that the variable in question is normally distributed with a

mean and standard deviation of the predictions of the CMIP5 set of models?

According to Winsberg (2018, 99):

> The average and standard deviation of a set of trials is only meaningful if those trials represent a random sample of independent draws from the relevant space- in this case the space of possible model structures. (Winsberg 2018, 99)

He further argues that this assumption is implausible since:

> What, after all, *is* the space of possible model structures? And why would we want to sample randomly from this? After all, we want our models to be as physically realistic as possible, not random. Perhaps we are meant to assume, instead, that the existing models are randomly distributed around the ideal model, in some kind of normal distribution. This would be an analogy to measurement theory. But modeling isn't measurement, and there is absolutely no reason to think this assumption holds. (Winsberg 2018, 99)

I agree with Winsberg in so far as the assumption that the set of available models represent a random sample from the relevant space of models seems to be a necessary assumption for the average and standard deviation of the multi-model ensemble results to be 'meaningful'. And I also agree that this assumption is extremely questionable for more than one reason. First, as Winsberg points out in the above passage, it is not at all clear what the space of possible model structures is even supposed to mean. Is there really a class of possible model structures of the climate system? What distinguishes a possible model structure from an impossible one? Is this class mathematically definable?[16] Without an answer to these questions, there is no reason to suppose that it makes sense to talk about the space of possible model structures in the first place. Second, the idea that the set of available models constitutes something like a random sample from this alleged space of possible model structures is extremely questionable.

[16] Frigg et al. (2014) argue that it is not at all clear how to circumscribe the class containing all possible model structures of a target system and raise doubts as to whether this class is mathematically definable.

For a start, why assume this in the first place? A simple random sample is a subset of a population in which each member of the subset has an equal probability of being chosen and statisticians usually go to great pain to make sure that the sample they take can indeed be reasonably thought to be a random sample from the population of interest. For instance, in the random number method, one assigns every individual a number and by using e.g. a random number generator, a subset of the population is randomly picked. If you have a hard time imagining the IPCC assigning numbers to all the models in the space of possible model structures and using a random number generator to pick a sample for it, you are not alone! But even if we were willing to (rather dramatically) stretch the statistical concept of a simple random sample, as Winsberg remarks:

> One obvious reason to doubt [this assumption] is that all of the climate models on the market have a shared history. Some of them share code; scientists move from one lab to another and bring ideas with them; some parts of climate models (though not physically principled) are from a common tool box of techniques, etc. (Winsberg 2018, 99)

Indeed, the IPCC authors themselves openly acknowledge that 'some features shared by many models are a result of the models making similar assumptions and simplifications' (IPCC 2010a, 10) and that 'different models may share components and choices of parameterizations of processes and may have been calibrated using the same data sets' (ibid., 6), concluding that 'models may not constitute wholly independent estimates' (ibid., 10). But then, if climate models are not constructed independently and are likely to share systematic sources of error, it really does seem unreasonable to assume that the CMIP5 set of models can be thought of representing anything like a random sample from the set of possible model structures.

But leaving aside considerations regarding the implausibility of this assumption, I'd like to further point out that this assumption alone does not explain why the IPCC authors are taking the 5-95% range of a normal distribution with the mean and the standard deviation of the multi-model ensemble results to derive

the interval [0.3°C, 1.7°C] as the *likely* range for the GMST81-100. In particular, where does the normal assumption come from?

As far as I can see, there are two routes one could take to attempt to justify where the normal assumption comes from. Under the first route the obtained interval is supposed to represent a confidence interval for the GMST81-100. Under the second route, the obtained interval is supposed to be a tolerance interval covering 90% of the results of the models in the space of possible model structures. Let me go through each of these routes (by the data provided by the IPCC, I am assuming that the mean of the results of the available models is equal to 1, the standard deviation of the models' results is 0.4 and the number of models available is 32).[17]

Under the first route, one would argue something like the following:

Argument 1

1. There is such a thing as the space of possible model structures.
2. The set of available models can be assumed to be a random sample from this space.
3. The mean μ of the results of all the models in the space of possible structures is the correct value for the GMST81-100.
4. By the Central Limit Theorem and premises 1, 2 and 3, the sampling distribution of the mean is assumed to be normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{N}}$, where σ is the standard deviation of the results of the models in the space of possible structures and N is the number of available models.

Conclusion The interval $[1 \pm 1.64\frac{\sigma}{\sqrt{32}}$ °C] is a 90% confidence interval for the correct value for the GMST81-100.

Unfortunately, although under this route it is clear where the normal assumption comes from, this is clearly not what the IPCC have in mind. For a start,

[17] the number of models available for each projection in the table above (Figure 2.8) is provided in Figure 12.5 of the AR5 (IPCC 2013, 1054).

notice that the argument above assumes that the standard deviation σ of the results of the models in the space of possible model structures is known. Whereas, the IPCC authors are using the standard deviation of the results of the available models. But the main reason to suspect that this is not what the IPCC have in mind is that the normal distribution they construct has the same standard deviation of the results of the available models, whereas if they had anything like a confidence interval in mind they would divide the sample's standard deviation by the square root of the number of models available (in this case 32). Hence this argument does not allow us to arrive at the same interval for the GMST81-100 as the one obtained by the IPCC authors and hence it cannot be the right route to justify where the normal assumption is coming from.

Under the second route, one would argue something like the following:

Argument 2

1. There is such a thing as the space of possible model structures.
2. The set of available models can be assumed to be a random sample from this space.
3. The results of the models in the space of possible model structures are normally distributed.
4. At least one result of a model from the space of possible model structures is the correct value for the GMST81-100.
5. It is reasonable to assume that the mean (=1) and standard deviation (=0.4) of the results of the available models are equal to the mean and standard deviation of the results of the models in the space of possible model structures.

Conclusion 1 The interval [$1 \pm 1.64 \times 0.4$°C]=[0.3°C, 1.7°C] is a (reasonable) tolerance interval covering 90% of the results of the models in the space of possible model structures.

Conclusion 2 It is reasonable to assume that there is a 90% probability that the correct value for for the GMST81-100 lies in the interval [0.3°C, 1.7°C].

This argument allows us to arrive at the same interval as the IPCC authors, so that's good news! However, aside from the controversial premise 2 already questioned above, there are a couple more premises in this argument that are extremely questionable too. For a start notice that premise 3 simply stipulates that the results of the models in the space of possible model structures are normally distributed. But it is unclear why one would assume such a thing. As mentioned in Section 2.5.1, in the problem of repeated measurement, it is often assumed that the parent population distribution is itself normal (again by appealing to the Central Limit theorem). But as Winsberg remarks 'modeling isn't measurement' so it is really unclear what would justify this assumption in this case. Premise 5 is also extremely questionable, for there is absolutely no reason to assume that the mean and standard deviation of the results of the available models are equal to the mean and standard deviation of the results of models in the space of possible model structures. Indeed, if we were to take seriously the idea that the interval obtained is supposed to be a tolerance interval covering 90% of the results of the models in the space of possible model structures then depending on the level of confidence we demand we would get a very different interval from the one obtained by the IPCC. For instance, a tolerance interval covering 90% the results of the models in the space of possible model structures at a 99% confidence level would be the interval [0.06 °C, 1.94°C], whereas the same tolerance interval but at a 95% confidence would be the interval [0.15 °C, 1.85°C]. The width of both of these intervals is substantially greater than the width of the IPCC interval [0.3°C, 1.7°C]. We could only obtain a tolerance interval covering 90% of the result of the models in the space of possible model structures equivalent to the interval obtained by the IPCC by demanding a mere 60% confidence level, which is a rather low level of confidence and certainly not one on which statisticians usually rely to make inferences.[18]

Overall, even if we were to grant the assumption that the available models represent something like a random sample from the space of possible model structures it is extremely unclear where the normal assumption comes from and

[18] I have once again relied on the statistical interactive page by Pezzullo (2005) to calculate these tolerance intervals.

more generally on what basis the IPCC authors arrive at the interval [0.3°C, 1.7°C] for GMST81-100. **Argument 2** seems to me the only plausible route one could take to justify why the IPCC arrive at this interval for GMST81-100 but as argued above this argument relies on multiple unwarranted assumptions, hence there is no reason to think this argument is sound.[19]

At this point someone might argue that perhaps we should not worry about all this after all. As mentioned at the beginning of this section, the IPCC authors are not claiming that the derived interval for the GMST81-100 is *very likely* as would follow from my second (unsound) argument. All they are claiming is that the correct value for GMST81-100 is *likely* (i.e. has at least a 66% probability) to be in that interval. And in the summary for policy makers it is explained that this downgrading of likelihood is due to an 'accounting for additional uncertainties or different levels of confidence in models' (IPCC 2013, 23). Hence, although a rigorous defense of the procedure that the IPCC authors use to arrive at the interval [0.3°C, 1.7°C] for GMST81-100 cannot be found, one may argue that one is not really needed: the IPCC approach is a rough, reasonable way to proceed, given all the extant challenges and uncertainties and it only gives us, at the end

[19]It is worth mentioning that according to the 'good practice guidance paper on assessing and combining multi-model climate projections' (IPCC 2010a, 4) 'statistical frameworks in published methods using ensembles to quantify uncertainty may assume (perhaps implicitly)':

a. 'that each ensemble member is sampled from a distribution centered around the truth ("truth plus error" view)[...] In this case, perfect independent models in an ensemble would be random draws from a distribution centered on observations.'

b. . 'that each of the members is considered to be "exchangeable" with the other members and with the real system [...] In this case, observations are viewed as a single random draw from an imagined distribution of the space of all possible but equally credible climate models and all possible outcomes of Earth's chaotic processes. A 'perfect' independent model in this case is also a random draw from the same distribution, and so is 'indistinguishable' from the observations in the statistical model.'

The practice guidance paper further writes that 'with the assumption of statistical model (a), uncertainties in predictions should tend to zero as more models are included, whereas with (b), we anticipate uncertainties to converge to a value related to the size of the distribution of all outcomes [...] While both approaches are common in published literature, the relationship between the method of ensemble generation and statistical model is rarely explicitly stated.' (IPCC 2010a, 4)

The IPCC multi-model ensemble methods discussed in this section may prima facie seem to fall under approach b, where the imagined distribution is for some reason assumed to be normal (an assumption that may, arguably, be inspired by approach a). However, as the IPCC guidance paper remarks although under this approach 'uncertainties *converge* to a value related to the size of the distribution of all outcomes' the IPCC does not use standard statistical concepts (such as tolerance intervals) which enable one to take account of the fact that what one has is merely a *sample* of the population, not the population itself. Hence, although as argued in this section both approach a and b rely on implausible assumptions, they are at least coherent. Whereas the IPCC 'multi-model ensemble methods' discussed in this section are in my view, incoherent.

of the day, something that is considered worthy of at least 66% of our credence. So maybe it's ok that it's not rigorously justified.[20]

However, I find this justification for not worrying about all this somewhat troubling, for more than one reason. First, the reasoning behind the downgrading of likelihood is very unclear and non-transparent. Why are the IPCC authors downgrading from *very likely* to *likely*? In other words, why is the range derived by these methods worthy of at least 66% of our credence, as opposed to say at least 60% or 55% of our credence? Given the lack of any detailed discussion on why the IPCC authors decided to declare this range *likely*, it is hard to resist the suspicion that the IPCC authors are simply downgrading likelihood to the next level down without any sort of serious expert deliberation, which is not a very reassuring thought. Second, if the IPCC authors themselves do not think that the normal distribution obtained by these "multi-model ensemble methods" really does represent the actual uncertainty for the variable in question, then it's not clear to me why they should rely on this distribution in the first place to determine a plausible the range for that variable. This is troubling because, despite the downgrading of likelihood in this case, the most salient feature of this finding is, in my view, the range [0.3°C, 1.7°C]; and given the substantial uncertainty in this range it seems to me highly misleading to draw so much attention to it. Related to this point, if the IPCC authors do not actually think this range is *very likely*, then it is not clear why they insist on reporting this very same range (but with a lower likelihood assignment). Intuitively another way to proceed would be to report a wider range instead, one for instance, that the authors consider *very likely* rather than just *likely*.

At this point one might object that it may be more epistemically demanding for the IPCC authors to provide a wider range that is considered to be *very likely* in contrast to a smaller range that is considered to be *likely*. That is, the IPCC authors may feel comfortable with saying that at least 66% of our credence should go in the range [0.3°C, 1.7°C] for GMST81-100, but they may not feel comfortable with saying that at least 90% of credence should go in a greater range. Hence,

[20] My thanks are owed to Wendy Parker for an email exchange in which she put forward some possible objections for me to consider.

we should not demand the IPCC authors to do something that they don't feel comfortable with doing. Furthermore, one might also object that my point about the range being the most salient feature of this finding is unjustified because all the IPCC authors are saying is that that range is worthy of at least 66% of our credence. Hence, the IPCC authors are not telling us we should believe that the correct value for GMST81-100 actually lies in that range. Hence again my worries are overall unjustified.

However, given the lack of any sort of rigorous justification for why the IPCC authors have decided to declare that the range [0.3°C, 1.7°C] is *likely* as opposed to another range (one with a greater/smaller lower limit or/and with a lower/greater upper limit) it is not clear why what the IPCC authors are doing would be evidently less epistemicallly demanding than what I am suggesting they might want to do instead.[21] Second, if indeed we are not supposed to take this range seriously (and hence not worry about the lack of a rigorous justification for how this range was derived in the first place) then this begs the question: what is the epistemic value of this finding? That is, if all the IPCC finding amounts to saying is that there is up to a 34% probability that the correct value for GMST81-100 lies outside the range [0.3°C, 1.7°C] without any direction as to how likely one should think that the correct value for the GMST81-100 is e.g. substantially higher than the upper limit of this range, then I just can't quite see on what basis a e.g. policy maker should regard the range [0.3°C, 1.7°C] to be the best estimate for the GMST81-100 on which to base their decisions. In other words, if the IPCC are not able to exclude the possibility that, for instance, the correct value for GMST81-100 may plausibly (with up to a 34% probability) be substantially higher than the range reported, then clearly a policy maker should know about this!

Together with Winsberg, I suspect that the multi-model ensemble methods used by the IPCC for quantifying uncertainty in their findings 'are grounded

[21] If the reason why what the IPCC authors are doing is less epistemicallly demanding than what I am suggesting they might want to do instead, is that what they are doing requires very little thinking on their part (i.e. using conceptually incoherent mechanical methods to derive a range for a variable and then simply call this range 'likely' because it sounds about right) then clearly this does not count as a valid reason.

and conceptualized out of a misguided desire to produce objective probabilities' (Winsberg 2018, 99).[22] This desire is misguided because it is essentially the desire to produce 'probabilities' that are independent of the individual beliefs of the IPCC authors, but at the cost of producing probabilities that are neither objective nor subjective. Hence it is not clear why we should call these 'probabilities' in the first place. The fact that in this case the IPCC authors are downgrading the likelihood level obtained from these methods from *very likely* to *likely* is, in my view, merely a cover up for the fact that the IPCC is not willing to step out of their comfort zone and embrace an 'approach that *self-consciously* reflect the subjective degrees of belief of the relevant set of experts' (Winsberg 2018, 100) in light of the models' results and their understanding of those models, their shortcomings, and the climate system. Of course, as I will discuss in Chapter 5, the question of how to interpret and communicate multi-model ensemble results is an incredibly hard challenge. However, this does not justify using mechanical-unjustified procedures for quantifying uncertainty to deal with this challenge.

Finally, recall that one of the reasons that I discussed in Section 2.4 for the emergence of two uncertainty scales in the AR4 uncertainty framework was the need to distinguish WG I and WG II distinct approaches in the characterization of uncertainty, which I interpreted as the need to distinguish subjective from frequentist/objective probabilities. However, in light of the multi-model methods discussed in this section, I fear that this reason may in fact be better interpreted as the need to make a distinction between subjective probabilities on the one hand, and neither subjective nor objective probabilities on the other. In other words, I fear that it may have to be interpreted as the desire to accommodate or even encourage WG I's misguided desire to produce probabilities in a mechanical way, at the cost of little, if any, objectivity at all. And if *this* was one of the reasons for the emergence of two uncertainty scales then it clearly was not a good one to begin with.

[22] By this I don't mean to suggest that there can't be a role for objective probabilities in climate science in general.

2.6 Taking stock

As seen in this section, the history of the IPCC calibrated uncertainty language is a convoluted one. Below is a diagram showing the history line of the IPCC uncertainty calibrated language.

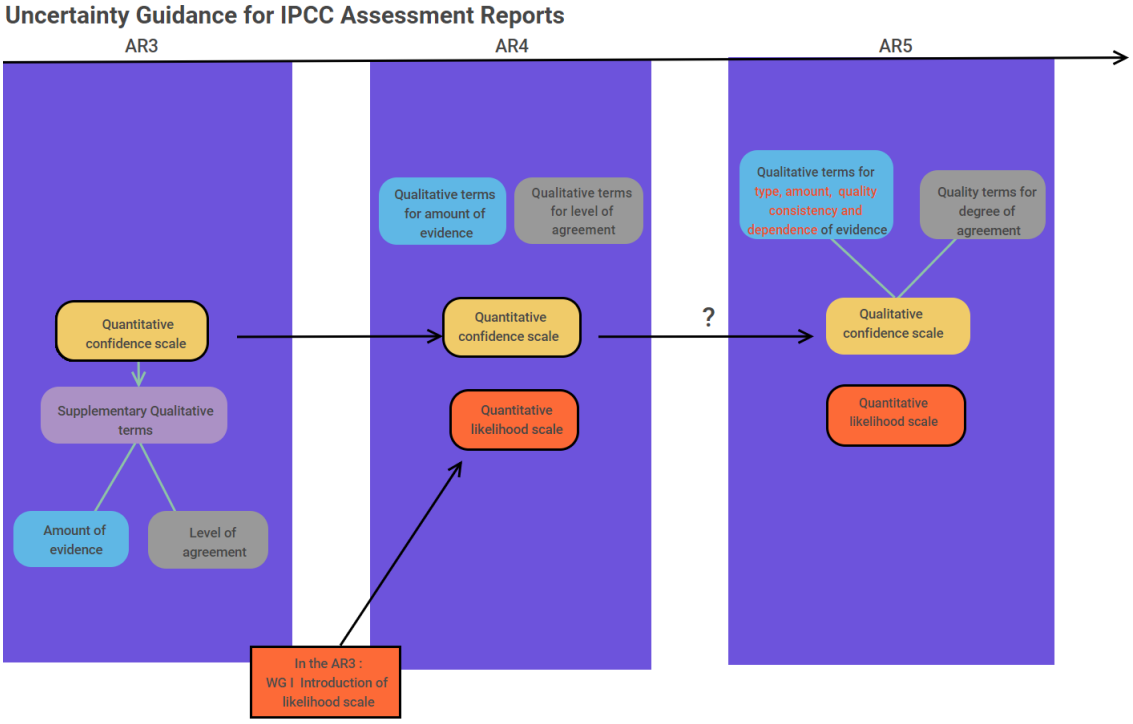

FIGURE 2.9: History line of the IPCC calibrated uncertainty language.

I have argued that the history of the IPCC calibrated uncertainty language can give us some insights into the problematic aspects of the AR5 uncertainty framework (IPCC, 2010) identified in Chapter 1. One issue, discussed in Section 1.3, was

1. *the puzzling bifurcation of evidence and agreement in the characterization of confidence.*

In Section 2.2, I have argued that the discrepant resemblance between the table of 'supplemental qualitative uncertainty terms' provided by the AR3 uncertainty guide (Figure 2.2) and the 'depiction of evidence and agreement statements and their relationship to confidence' that one finds in the AR5 uncertainty guide (Figure 1.1) shows that the puzzling bifurcation of evidence and agreement in

the characterization of confidence is partly due to an ill-thought-out stopgap that first appeared in the AR3 uncertainty guide and that got petrified in the process, despite its irreconcilability with the revisions to the IPCC uncertainty guides later introduced. If this is right, then there really doesn't seem to be an epistemically good reason for the bifurcation of evidence and agreement in characterization of confidence, and hence should, arguably, be removed.

The history of the IPCC calibrated uncertainty language also teaches us something about the issue discussed in Section 1.4:

2. *The lack of an adequate interpretation of confidence and likelihood compatible with the AR5 uncertainty guide's recommendations*.

In Section 2.4, I have argued that an understanding of the *reasons* behind the first appearance of two uncertainty scales ('confidence' and 'likelihood') in the AR4 uncertainty framework can give us some insights into the nature of this issue. I have argued that there were two distinct reasons. On the one hand, the emergence of two uncertainty scales was due to a felt need to distinguish WG II approach in the characterization of uncertainty, which focused on the degree of understanding and consensus from that of WG I, which focused on frequentist statistics instead (**reason 1**). On the other hand, this emergence was also due to a felt need to distinguish the assessed probability of an outcome (through the use of frequentist statistics), and the confidence that the science community had in its ability to determine it (**reason 2**). I have argued that there is a clear tension arising from these two distinct reasons behind the emergence of two uncertainty scales. If the IPCC authors lack confidence in the assumptions that justify the use of a frequentist approach in the characterization of uncertainty, then it is no longer clear why one should rely on a frequentist approach in the first place, which in turn means it is no longer clear how one should *interpret* those assessed probabilities: those assessed probabilities are neither *subjective* nor *objective*! I have further suggested that both the AR5 guide's recommendation to only use 'likelihood' in cases where confidence is sufficiently high is an attempt to deal with this tension. The underlying thought behind these recommendation is, arguably, something like this: if the assessed probability is to be

'objective' then likelihood should only be used if there is sufficiently high confidence in the assumptions that justify the use of a frequentist approach. But this is evidently not an adequate attempt to deal with this tension. Why? Because it is simply an attempt to *cover up* the apparent tension that arises from these two distinct reasons behind the emergence of two uncertainty scales, while (as argued in Section 1.4) providing recommendations that are incompatible with any possible interpretation of 'likelihood' and 'confidence' and hence making it impossible, for any one who tries, to understand what kind of uncertainty the likelihood and confidence metrics are actually supposed to represent.

In Section 2.5, I have further argued that some common frequentist approaches (i.e.'multi-model ensemble methods') on which the IPCC authors rely to assess uncertainty in their findings are not conceptually coherent methods for producing objective probabilities. Hence, these methods produce 'probabilities' that are neither objective nor subjective. Given this, I have argued that neither should these methods be relied on for assigning a likelihood level to an event, nor should they play a role in determining the range of 'plausible values' (i.e. what the IPCC call a *likely* range of values) for a quantity of interest. Finally, I have argued that the fact that these common frequentist approaches in the characterization of uncertainty are not conceptually coherent methods for producing objective probabilities raises some doubts as to whether what I called **reason 1** for the emergence of two uncertainty scales (i.e. the need to distinguish subjective approaches from frequent approaches in the characterization of uncertainty) was in fact a good reason to begin with. For if this reason cannot be interpreted as the need to distinguish subjective from objective probabilities, then it may have to be interpreted as the need to accommodate or even encourage WG I's misguided desire to produce 'probabilities' in a mechanical way, at the cost of little, if any, objectivity at all. And if *this* is one of the reasons for the emergence of an additional likelihood scale alongside the confidence scale in the IPCC uncertainty framework, then it is evidently a bad one.

The question of how uncertainty should be conceptualized by the IPCC is

undeniably a hard one, and I have sincere respect for all those who have attempted to address it. But overall, what I think the history of the IPCC uncertainty framework really teaches us is that any attempt to revise and improve the IPCC uncertainty framework must start from the recognition that for an uncertainty framework to be adequate, all key concepts involved in this framework must be unambiguously defined. When concepts such as probability, likelihood, confidence, agreement, and robust evidence lack an unambiguous interpretation, they are bound to be misunderstood, misused, even abused. Hence it is clear that any attempt to address the conceptual issues that I have discussed in Chapter 1 must start with the clear-sighted recognition that these issues will not be suitably tackled without first providing a clear interpretation of all the concepts involved.

There are several recent proposals for a new IPCC uncertainty framework that significantly depart from the current one, which I will critically assess in Chapter 7. However, I think that any sincere and successful attempt to revise and improve the current IPCC uncertainty framework will require a clear diagnosis of the conceptual problems it currently faces and why these have developed. In these two chapters, I have sought to contribute to this diagnosis.

Part II

Model-based robustness analysis

Chapter 3

Robustness analysis as tool for discovering robust theorems

3.1 Introduction

Any model of a real world phenomenon is bound to include idealizations of some sort (by disregarding some variables, or ignoring or simplifying interactions amongst variables, etc.). Yet we use models to learn about the world constantly, and shall not cease doing so any time soon. A question thus arises: why can we use models to learn about the world despite their idealizing assumptions? If no model is ever a complete and veridical representation of its target system, why do we think of them as 'vehicles for learning about the world' (Frigg and Hartmann, 2020)?

There is an idea pertinent to this question that is popular amongst some scientists and philosophers (e.g. Levins, 1966; Weisberg and Reisman, 2008; Kuorikoski et al., 2010, Schupbach, 2018). This idea broadly consists in the following: we can increase our confidence in a model's conclusion by 'studying a number of similar but distinct models of the same phenomena' (Weisberg 2013, 156). Learning that all these models give the same conclusion, it is claimed, should make us more confident in that conclusion. This way of dealing with model results is usually referred to by its proponents as 'robustness analysis'.

The first explicit discussion of robustness analysis in the context of modelling is usually attributed to the scientist Richard Levins (1966). Below is a frequently quoted passage from Levins on the notion of robustness:

> Even the most flexible models have artificial assumptions. There is always room for doubt as to whether a result depends on the essentials of a model or on the details of the simplifying assumptions. [...] Therefore, we attempt to treat the same problem with several alternative models each with different simplifications but with a common biological assumption. Then, if these models, despite their different assumptions, lead to similar results we have what we can call a robust theorem which is relatively free of the details of the model. Hence our truth is the intersection of independent lies. (Levins 1966, 423)

Levins' suggestive remark that 'our truth is the intersection of independent lies' became something of a shibboleth for advocates of robustness analysis. However, what this shibboleth means, and whether or not there is any truth in it, is to this day a source of contention. On the one hand, there are those who argue that robustness analysis has a rightful claim as a method of confirmation (e.g. Weisberg 2006, 2013;[1] Lloyd (2010); Kuorikoski et al. 2010, 2012); and on the other, there are those who disagree (e.g. Cartwright 1991; Orzack and Sober 1993; Odenbaugh and Alexandrova 2011, Justus 2012).

Despite the fact that advocates of robustness analysis as an epistemic tool disagree about what argument one should rely on to support this idea, they do seem to share at least one conviction: that the epistemic import of robustness analysis comes from its ability to distinguish results that are *artefacts* of a model's idealizations/assumptions from those that are not. Unfortunately, however, what this conviction *actually* consists in is, in my view, far from clear. Consider, for instance, the following quotations about some empirical results found in a variety of scientific journals:

> Fewer dolphins were sighted and acoustically detected after days of big swells. That the same pattern was evident in both the visual and acoustic data sets indicates that it *is not an artefact* of poor sighting

[1] As I will discuss in this chapter, Weisberg might not belong to this camp after all since according to him '[r]obustness analysis helps to identify robust theorems, but it does not confirm them.' (Weisberg 2006, 742).

> conditions on days with significant swell. (Dittmann et al. 2016, my emphasis)
>
> Together, these data suggest that inactivation is a function of the calcium channels and *was not an artefact* of contamination by some other current. (Schnee and Ricci 2003, my emphasis)
>
> Attempts at disguising the addition of energy in the HED version has been successful and [...] establish whether the reported changes in liking *were not an artefact* of unexpected changes in sensory quality. (Gould 2009, my emphasis)

I understand what is being asserted in these quotations: these are *causal* claims. The sight of fewer dolphins is not an artefact of poor sighting if poor sighting *did not cause* the sight of fewer dolphins. Inactivation is not an artefact of contamination by some other currents if contamination of some other currents *did not cause* inactivation. Reported changes in liking were not an artefact of unexpected changes in sensory quality if unexpected changes in sensory quality *did not cause* the reported changes in liking. They also seem epistemically valuable claims; what we learn from them is that a particular result/observation is not due to a specific cause unrelated to the subject of investigation (and that might thus lead to an incorrect conclusion in that investigation).

So I understand what is at stake in those quotations. By contrast, consider the following:

> In economics, proving robust theorems from different models with diverse unrealistic assumptions helps us to evaluate which results correspond to important economic phenomena and which ones *are merely artefacts* of particular tractability assumptions. (Kuorikoski et al. 2010, my emphasis)
>
> To determine whether a theoretical result actually depends on core features of the models and *is not an artifact* of simplifying assumptions, theorists have developed the technique of robustness analysis,

> the examination of multiple models looking for common predictions. (Weisberg and Reisman 2008, my emphasis)
>
> [Robustness analysis] allows us to learn if a model's result *is merely an artefact* of an idealization or if it is connected to a core feature of the model. (Weisberg 2013, my emphasis)
>
> But another possible explanation might be that the calculated result *is an artifact* of the large grid size of the simulation. (Winsberg 2018, my emphasis)
>
> [...] we may discard worries that our result *is an artefact* of a particular unrealistic assumption of the first model by using a second model that does not share that assumption. (Schupbach 2018, my emphasis)

These quotations, I find, are distinctly less clear. For although it seems evident the authors of these lines share the implicit assumption that it should be obvious what it means to claim that *a model's result is not an artefact of an idealization*, I don't think it is. What *could* it mean? For a start, notice that this cannot be a claim about whether or not an idealization of the model played a role in the derivation of the model's result, since *all* of the model's assumptions and idealizations were used in the derivation of a model's result. In other words, in contrast with the previous claims, this cannot be a claim about what *caused* a model's result. It is also hard to see how this could be a claim about whether or not the model's result is true of the model's target system. Firstly, the fact that a model's result is not an artefact of one particular idealization does not tell us whether or not it is an artefact of some other idealization; learning that a model's result is not an artefact of a particular idealization could not therefore, it seems to me, constitute a direct claim about a real-world phenomenon. If it did, why claim a model's result is not an artefact of a particular idealization, rather than another? Secondly, it seems that the result could be true of the target system *despite* the model's result being an artefact of its idealizations (by mere coincidence!).

Although the claim that a model's result is not an artefact of an idealization cannot be understood as direct claim about a real world phenomenon, perhaps

it could be understood as a *relational* claim instead, in particular a claim about how the idealization in question *relates* to the real world phenomenon under investigation. One way to cash this out could be to interpret the claim that a result is not an artefact of an idealization as the claim that the idealization in question merely distorts features of the target system that are not causally relevant to the phenomenon under investigation. Indeed, as Rice (2012, 183) remarks 'most accounts of mechanistic modelling assume that successful models (for example, those that can explain) are those models whose components and interactions accurately represent the relevant (that is, difference-making) causal relationships among the components of the target mechanism(s) and leave out—that is, abstract away—irrelevant features'. But then could the authors of the above quotes have something like this in mind? In other words, should we interpret the claim that a result is not an artefact of idealization in the above quotes as the claim that the idealization is 'harmless' (i.e. it merely distorts features of the target system that are not causally relevant to the phenomenon under investigation)?

Leaving aside the various (serious) challenges that Rice raises against these accounts of mechanistic modelling,[2] it seems to me that the authors of the above quotes cannot have something like this in mind either. This is because if all it takes to discard worries that a model's result is an artefact of an idealization is to learn that another model which does not share that particular idealization gives that same result, as all the authors of these quotes seem to implicitly or explicitly assume, then the claim that a model's result is not an artefact of a idealization cannot be understood as the claim that that idealization is 'harmless' in the above sense. This is because this would entail that one would be entitled to infer that an idealization is 'harmless' (i.e. that the feature of the target system distorted by the idealization is causally irrelevant to the phenomenon under investigation) from the mere discovery that some models which do not share a

[2]One important objection that Rice (2019) raises against these accounts is that the assumption that 'models can be decomposed into the contributions made by their accurate and inaccurate parts' (ibid., 180) on which these accounts of mechanistic modelling rely, is often not plausible since in many cases of modelling 'idealizations are not innocent bystanders that can be quarantined by only distorting irrelevant (or insignificant) features; instead, they are deeply invested collaborators that allow for the application of various mathematical modelling techniques' (ibid., 194).

particular idealization happen to give the same result. And this is not plausible for at least two reasons. First, inferring a direct claim about the target phenomenon (i.e. that a particular feature is causally irrelevant to the phenomenon under investigation) from looking at the behaviour of some models would seem to require a justification for why those models are relevant for inferring such a thing. But to the best of my knowledge the authors of the above quotes offer no such justification. This is, arguably, evidence that this cannot be what they have in mind; 2) the models involve *several* idealizations not just the one in question. And without knowing whether or not the models in question may involve *other not* 'harmless' idealizations, it's very unclear why looking at the behaviour of these models could allow us to automatically infer a direct claim about a real world phenomenon in the first place.

In light of the above discussion, it seems to me the only plausible interpretation of the claim a model's result is not an artefact of an idealization in the above quotations must concern the properties of the model relative to another model. But *which* other model is far from obvious. Below are some suggestions.

My first suggestion is the following:

Interpretation I.1. *A model's result R is not an artefact of idealization A_1 if when replacing A_1 with a more realistic assumption the (new) model would give the same result R.* [3]

Under this interpretation, the claim that a model's result R is not an artefact of idealization A_1 is merely a claim about a property of the model in relation to another model. Hence on its own it tells one nothing about whether or not the model's result is actually true of the target system. To see this, suppose the model contains another idealization B_1. Clearly the fact that the claim is true on its own entails nothing about whether R is in fact true. All we know is that $A_1 + B_1+$ (all the other model's assumptions and idealizations) entail R and that replacing A_1 with a more realistic assumption does not affect this result. But this says nothing about whether or not R is an artefact of e.g. B_1; and if this were the case, then there would

[3] When an idealization is replaced with a more realistic assumption, this is usually described in the literature as the de-idealization of a model.

be no reason to think that R is actually true of the target system. Notice, further, that a necessary condition for the claim to be true under this interpretation is that the idealizing assumption A_1 *can* be replaced with a more realistic assumption in the first place. These types of idealizations are known in the literature as 'Galilean idealizations' (McMullin, 1985).

Here is another suggestion:

Interpretation I.2 *A model's result R is not an artefact of idealization A_1 if when replacing A_1 with a weaker assumption, i.e. an assumption that is entailed by A_1, the (new) model would give the same result R.*

This is a rather different interpretation. Here, what one learns from the claim that a model's result R is not an artefact of idealization A_1, is that R can also be derived from replacing A_1 with a weaker assumption, rather than a more realistic one (as under **interpretation I.1**). And although a weaker assumption can often be a more realistic one too, this is not always the case (e.g. the assumption that 'my neighbour's cat can speak Italian and/or it is black' is weaker but in no sense more realistic than the assumption that 'my neighbour's cat is black'). From a mathematical perspective, this is certainly an interesting claim to learn for as Raz (2017) points out, learning that a result can be generalised seems to be relevant to the mathematical explanation of a result. Clearly, however, as Raz points out 'the generalization of a result is only as good as the idealizations used in the general model' (Raz 2017, 751).

Here is a final suggestion:

Interpretation I.3 *A model's result R is not an artefact of idealization A_1 if when replacing A_1 with another (equally unrealistic) idealization the (new) model would give the same result R.*

From a mathematical point of view, learning that R can be derived by replacing A_1 with another idealization A_2 does not strike me as a particularly interesting fact in itself: it might for all we know be a mere mathematical

coincidence.[4] Of course, learning this fact might prompt one to search for a mathematical explanation for this apparent coincidence (for instance, by showing that these two distinct models are special cases of a more general model which also gives R); and one might find it! But whether or not one finds it, and whatever mathematically valuable 'discovery' one may make along the way, it is not the prompting fact that is mathematically (and hence perhaps also epistemically) valuable, but that subsequent development.

It should, of course, be mentioned that idealizations are not the only source of worries when it comes to modelling. One might often also lack trust in a model's result because the model involves assumptions about the target system that *could* be true or false (in contrast with idealizations); assumptions one is *unsure* are true. What if the claim was not about an idealization, but about an assumption? That is, what could the claim that 'a model's result is not an artefact of a model's assumption' mean? Here are three suggestions, in spirit akin to those that came before:

Interpretation A.1. *A model's result R is not an artefact of assumption A_1 if when replaced with a true assumption the (new)*[5] *model would give the same result R.*

Under this interpretation when we learn that a model's result R is not an artefact of a particular assumption A_1, we learn that if A_1 were to be replaced with a true assumption A_k, the (new) model would give the same result R. Notice, however, that we could learn this fact despite not knowing what assumption A_k actually consists in. For instance, assume that there is a finite set of possible assumptions $A_2, A_3, \ldots A_n$ about the target system that can replace A_1 and that we know that one of them must be right, despite not knowing *which one* is right. Suppose further that we learn that result R can be derived from the set of models consisting of the

[4] I am using the notion of mathematical coincidence in the sense of Lange (2010). According to him it is a coincidence that two mathematical facts are true iff they have no 'single unified explanation' i.e. a proof that explains 'why (and prove[s] that) all of the components of the non-coincidence are true if any one is true - that is, why they all stand or fall together'. (Lange 2010, 327)

[5] Strictly speaking the new assumption could be same assumption as A_1 so in this case using the word 'new' would be inappropriate.

model we started with (which involves A_1) and the models for which A_1 has been replaced with $A_2, A_3, \ldots$ and A_n respectively. In this case we learn that when replacing assumption A_1 with a true assumption A_k for some $k \in (1, 2, \ldots, n)$, the (new) model gives the same result R, despite not knowing the value of k. As all the other interpretations, under this interpretation the claim that a model's result is not an artefact of assumption A_1 is first and foremost a claim about a property of the model in relation to another.

Interpretation A.2. *A model's result R is not an artefact of assumption A_1 if when replacing A_1 with a weaker assumption, i.e. an assumption that is entailed by A_1, the (new) model would give the same result R.*

Very similar considerations to the ones I made about **interpretation I.2** apply to this interpretation.

Interpretation A.3. *A model's result R is not an artefact of assumption A_1 if when replaced with another not necessarily true assumption the (new) model would give the same result R.*

Notice that in some special cases learning that a model's result R is not an artefact of assumption A_1 under this interpretation entails that we also learn this claim under **interpretation A.1**. Suppose, for instance, that there is only one possible assumption A_2 that can replace A_1, and that we know that A_k is true for some $k \in [1, 2]$. Clearly in this case, if we learn that a model's result R is not an artefact of assumption A_1 under this interpretation we also learn the same claim under **interpretation A.1**.

This is not at all meant to be an exhaustive list of all possible interpretations of the claim that a model's result is not an artefact of an idealization/assumption. However, by offering these distinct possible interpretations, I want to stress the fact that an assessment of the epistemic value of learning that a model's result is not an artefact of an idealization/assumption requires first and foremost an understanding of what such learning consists in. Unfortunately, in my view, proponents of robustness analysis as an epistemic tool are often not sufficiently clear

about what such learning does consist in. This lack of clarity is, I believe, *partly* responsible for the confusion and contention surrounding the epistemic import of model-based robustness analysis. Clearly, however, it is impossible, in the abstract, to settle as much; hence only an investigation of actual cases of robustness analysis can help us understand what its proponents might implicitly mean by the claim that a model's result is not an artefact of an idealization/assumption.

The overall aim of this chapter is to critically assess the view that robustness analysis has a rightful claim as a method of *discovery* of what are known in the literature as "robust theorems", which are theorems of the general form: 'Ceteris paribus, if [common causal structure] obtains, then [robust property] will obtain' (Weisberg 2006, 738).[6] Its structure is as follows. In Section 3.2, I will discuss the discovery of the Volterra principle through the analysis of predator-prey models, a principle which is considered 'an especially striking example of a "robust theorem"' (Weisberg and Reisman, 2008) in the literature on robustness analysis. Through this example, I will show that the claim that a model's result is not an artefact of an idealization/assumption must often be used with various different interpretations implicitly in mind. In Section 3.3, I will discuss in detail Weisberg's general characterization of robustness analysis. I will argue that by accepting that 'low-level confirmation' automatically confirms 'robust theorems', as Weisberg does, one must at the same time accept that robust theorems do not have to concern the actual world for them to deserve the name and hence that one is not warranted to assume that they can be useful for explaining or predicting real-world phenomena, contrary to what is usually assumed in the literature. Hence, I will conclude that if one thinks that robust theorems must concern the actual world, one cannot assume that they are automatically confirmed by Weisberg's notion of low-level confirmation.

3.2 Robustness reasoning "in action"

The Lotka-Volterra model (independently proposed by Volterra (1926) and Lotka (1956)) is used to represent the behaviour of real-world predator-prey systems

[6] Whereas in the next chapter, I will critically assess the view that robustness analysis has a rightful claim as a method of *confirmation* of "robust theorems".

and is described by the following two coupled ordinary differential equations:

$$\frac{dV}{dt} = rV - (aV)P \tag{3.1}$$

$$\frac{dP}{dt} = b(aV)P - mP \tag{3.2}$$

Where $V(t)$ and $P(t)$ stand for the size of the prey and predator population at time t, respectively. The constant r stands for the birth rate of the prey population and the constant m stands for the death rate of the predator population. The constant a stands for the predator attack rate and the constant b stands for the predator conversion efficiency.

The Lotka-Volterra model involves several idealizing assumptions. For instance, the model assumes that prey are born at a single constant rate, or that predators have no saturation, that is that their consumption rate is potentially unlimited. It assumes that the predator attack rate is not affected by the size of the predator population nor by any other plausible factors (such as the number of refuges the prey have access to and many others). We know that these are false simplifying assumptions that do not hold for any *real-world* predator-prey system. So in light of these idealizations, it is not clear why one should think of the Lotka-Volterra model as an adequate representation of real world predator-prey systems and hence, due to this, why we should trust any of its results to hold in real-world predator-prey systems.

However, there is one result in particular that is of interest to Weisberg and Reisman (2008). A straight forward calculation reveals that the ratio of the equilibrium value of the predator population, $\widehat{P}$, to that of the prey population, $\widehat{V}$ is given by

$$\rho = rb/m. \tag{3.3}$$

And an investigation of this equation reveals that the introduction of an external factor that decreases the prey growth rate, r, and increases the predator death rate, m, will decrease the value of ρ. But the equilibrium values for P and V are

also the time averages for the predator and prey populations respectively in this case.[7] Hence this shows that the introduction of an external factor that decreases the prey growth rate, r, and increase the predator death rate, m, will decrease the ratio of the time averages for the predator population, $\bar{P}$, to the time average of the prey population $\bar{V}$. This result, known as the *Volterra Property*, is interpreted by Weisberg and Reisman (2008, 113) as follows:

> The introduction of any substance that has a harmful effect on both predators and prey (a general biocide), will increase the relative abundance of the prey population.

Weisberg and Reisman show that what is special about the Volterra property is that it is present across several other models that, despite being different in several respects, all share a common assumption: the predator-prey system is negatively coupled, i.e. 'increasing the abundance of predators decreases the abundance of prey and increasing the abundance of prey increases the abundance of predators' (ibid., 114). This, according to them, shows that the following principle, which is an example of what Weisberg calls a 'robust theorem', is a true empirical hypothesis.

> *The Volterra principle*: Ceteris paribus, if a two-species, predator-prey system is negatively coupled, then a general biocide will increase the abundance of the prey and decrease the abundance of predators.

In Section 3.3, I will look in detail into why according to Weisberg we should believe the Volterra Principle (and robust theorems in general) to be a true empirical hypothesis. But before I do that let us first have a look at some of the models they consider to conclude that the Volterra principle is a "robust theorem" in the first place.

To demonstrate that the Volterra principle is 'an especially striking example of a robust theorem', Weisberg and Reisman (2008) make an important distinction between parameter robustness analysis, structural robustness analysis, and

[7] Due to other properties of the Lotka-Volterra model in this case the average abundance of a system does coincide with the equilibrium. However, the existence of an equilibrium does not in general imply that the average abundance of a system will coincide with that equilibrium.

representational robustness analysis and argue that the Volterra property is robust under each of these three distinct kinds of robustness analysis. According to them,

> Taken together, these three kinds of robustness analysis are a powerful way of demonstrating that a particular modeling result is not dependent on the particular assumptions or idealization embodied in a model or family of models. (Weisberg and Reisman 2008, 108)

As I will show in this section, if each of these distinct kinds of robustness analysis (parameter, structural, and representational) is a way of demonstrating that a particular model's result is not an artefact of particular assumptions or idealizations, then this must mean that Weisberg and Reisman have a rather liberal conception of the claim that 'a model's result is not an artefact of an idealization or an assumption' (i.e. they must embrace various different interpretations of this claim at once). For, as we will see, each kind of robustness has an associated distinct class of models for which the Volterra property is derivable; and hence each kind of robustness involves showing that the Lotka-Volterra property is not an artefact of an idealization/assumption under different interpretations. Let us look at each in turn.

Parameter robustness analysis involves checking whether a model described by the same equations as the original Lotka-Volterra model but with different parameter values gives the same result (i.e. the Volterra property). According to Weisberg and Reisman, all parameter values where the two species coexist yield the Volterra property. So the Volterra property seems to exhibit parameter robustness. *But what fact do we learn from this?* That clearly depends on what one thinks of the Lotka-Volterra model in the first place. If one believes that the Lotka Volterra's model is an accurate representation of the phenomenon of interest then the fact that the Volterra principle is robust across all parameters shows that despite the fact that one might not be sure about what the correct parameter values are, the same result will be derived by a model with the correct parameter values (regardless of what those might be). Hence what one learns

from this fact in this case is that *a model's result R is not an artefact of an assumption* under **interpretation A.1**.[8]

On the other hand, if one does *not* think that the Lotka-Volterra model is an accurate representation of the phenomenon of interest then it is not so clear what fact we are learning from parameter robustness. This is because some of the parameters might not have a straightforward physical interpretation in this case. To see why this may be, suppose that we don't think that the assumption that the prey birth rate is constant is a reasonable assumption to make about a particular real predator-prey system since we have good reasons to believe that the amount of available resources in the environment (which affects the birth rate) will greatly vary over time due to e.g. the changing size of the prey population over time or perhaps other factors (such as a fluctuating environment which may affect the amount of available resources).[9] In this case a specific value for the constant r can no longer be assumed to be a factual assumption about what is the prey birth-rate because the birth rate is not a constant! Hence, it seems to me that the right way to interpret a specific choice of a parameter value for the prey birth rate r, in this case, is not as an assumption that could be true or false about the system, but as an idealization that is known to be false no matter what parameter value we pick. Hence what one seems to be learning from parameter robustness in this case is that *a model's result R is not an artefact of an idealization* under **interpretation I.3** (and perhaps also **interpretation I.1**, as long as we believe that some parameter values for the prey birth rate, although still idealization, are in some sense more realistic than others).

Structural robustness analysis involves making structural changes to the Lotka-Volterra model, while keeping the core negative coupling intact. An example of a structural change Weisberg and Reisman consider is the addition of a maximum carrying capacity to the growth rate of the prey (i.e. in the absence of predators, the prey population is no longer assumed to grow exponentially as

[8]Refer to Section 3.1 for a description of this interpretation and for all the others that I will mention in this section.

[9]Indeed virtually all biological populations live in a seasonal environment, but the strength of the seasonality varies enormously, as does an organism's response to it. See, for instance, Vandermeer (1996) and Sauve et al. (2020) for examples of how one may modify the Lotka-Volterra model so to make the prey birth rate or the predator attack rate dependent on seasonal fluctuations.

there is now a maximum size to which it can grow). This is achieved by making the prey population growth rate density dependent so that the new model is described by the following two equations:

$$\frac{dV}{dt} = r(1 - \frac{V}{K})V - (aV)P \tag{3.4}$$

$$\frac{dP}{dt} = b(aV)P - mP \tag{3.5}$$

They go on to show that this model also exhibits the Volterra property and hence that this is in an instance of structural robustness. But what do we learn from this fact? It seems to me that what we are learning in this instance is that by replacing an idealization in the original Lotka-Volterra model (i.e. the assumption that the prey population will grow exponentially in the absence of any predator) with a more realistic assumption (i.e. that there is a maximum carrying capacity to the growth rate) the (new) model will give the same result (i.e. the Volterra property). But then this seems to be a case where we learn that *the model's result is not an artefact of a model's idealization* under **interpretation I.1**.

According to Weisberg and Reisman:

> Further structural robustness analysis would consider other changes to the causal structure represented in the model drawn from the kinds of ecological factors known to be relevant to population dynamics and predation. *While any change to the basic structure is a kind of structural robustness test*, ecologists are most interested in the ones that are potentially ecologically realizable. When a robust property survives all or some range of structural robustness tests, then we can say that the property is structurally robust to such and such changes to the causal structure of the system. *If these changes sample a sufficiently broad set of ecologically plausible circumstances*, then ecologists will often simply refer to a phenomenon as robust. (Weisberg and Reisman 2008, 119, my emphasis)

The emphasized remarks in this quote suggest that according to them structural robustness analysis can also allow us to learn that model's result is not an artefact of an idealization or an assumption under various other interpretations (such as **interpretation I.3** or **interpretation A.1**).

It is worth pointing out that Raz (2017) demonstrates that as long as a condition that ensures that the average abundance of a system coincides with the relevant equilibrium is satisfied (see Raz 2017, 748), the Volterra principle holds for a more general model[10] described by the following coupled ordinary differential equations:

$$\frac{dV}{dt} = rf(V)V - p(V)P \tag{3.6}$$

$$\frac{dP}{dt} = p(V)P - mP \tag{3.7}$$

where $f(V)$ and $p(V)$ are assumed to be differentiable for $V \geq 0$ with $\frac{df}{dt} \leq 0$ and $\frac{dp}{dt} > 0$. And $f(0) = a > 0$, $p(0) = 0$.

Since this model encompasses both the original Lotka-Volterra model and the new model considered above, 'this generalization shows that the models investigated by Weisberg are not really independent, but rather belong to the same type, and that they all satisfy the Volterra Principle, because they are of this type' (Raz 2017, 751). So from Raz's analysis one also learns that *the model's result R is not an artefact of an idealization* under **interpretation I.2**.

Finally, *representational robustness* analysis involves changing the 'representational framework' of the Lotka-Volterra model and assessing whether the same result (i.e. the Volterra property) still obtains. The following quote clarifies what Weisberg and Reisman mean by the representational framework of a model:

> Mathematical models can be thought of as being composed of state variables, which are variables that represent the properties (states) of interest to the modeler and transition rules, the rules that govern how the states change through time. [...] The representational framework of the model is a general description of the type of state

[10] Which is a slight modification of one proposed by Gause (1934).

> variables and the type of transition rules the model employs. For example, the variables in a biological model might represent individuals or populations. [...] Transition rules can be deterministic, probabilistic, or stochastic. They can also be discrete or continuous with respect to time. (Weisberg and Reisman 2008, 120)

Indeed there are always several possible mathematical frameworks to choose from when modelling any phenomenon. For instance, the Lotka-Volterra's model uses population state-variables, whereas one could choose to model a prey-predator system using individual state variables instead. It also has deterministic transition rules that are continuous with respect to time. But one could very well choose to make them discrete with respect to time etc.

As an instance of representational robustness analysis, Weisberg and Reisman demonstrate that the Volterra property can be derived[11] using a (density-dependent) individual-based model in which the Lotka Volterra model's variables, parameters and other assumptions are all translated into individual-based terms and which also defines a negatively coupled predator-prey system.[12] Hence we find that the Volterra Principle also holds in this model. *But what do we learn from this fact?* Are we learning that the Lotka- Volterra property is not an artefact of a particular idealization? And if so under what interpretation?

As Lisciandra (2017) points out, one clear difference between the population-based and the individual-based Lotka Volterra model is that the former assumes that the population is continuous whereas the latter assumes that the population is discrete, which is evidently a more realistic assumption. Could then this be an instance in which we learn that a model's result is not an artefact of an idealizing assumption under **interpretation I.1**? No, since as Lisciandra (2017) remarks:

> On the one hand, the fact that an individual-based model which is based on discrete populations gives the same result as the Lotka-Volterra model is an indication that the Volterra principle can also be

[11] This is done via the investigation of computational simulations rather than mathematical analysis.

[12] In this case a model that defines a negatively coupled predator-prey system is one for which (Ceteris Paribus) increasing the abundance of predators decreases the expected number of prey and increasing the abundance of prey increases the expected number of predators.

> derived under the assumption of discrete populations. On the other hand, however, when translating the Lotka-Volterra model into an individual-based model, many aspects of the initial model change. These changes come within an entirely new modeling 'package', whose assumptions will have to be tested in turn. Note that the more aspects have been changed, the further we are from analyzing the effect of one specific assumption. (Lisciandra 2017, 83)

Indeed, the population-based model and the individual-based model differ in all sorts of idealizations (for instance the individual based model involves several idealizations with respect to the behavioural rules of the individuals and the spatial representation of their environment whereas the population based Lotka-Volterra model clearly does not), not just in the assumption the population is assumed to be discrete rather than continuous. Hence, on further thought it does not seem that one can assert that the fact that the Volterra property obtains in this new representational framework shows that the Volterra property is not an artefact of a particular idealization under any interpretation discussed in Section 3.1. But if we don't learn that the model's result is not an artefact of an idealization, then what do we learn? I think this case should be thought of as a mathematical coincidence (in the sense of Lange (2010))[13] of questionable epistemic significance in and of itself. Suppose, for instance, that one were to construct a model with a different representational framework which defined a negatively coupled predator-prey system but that did not manifest the Volterra property. What would we learn in this case as far the Volterra principle is concerned? Clearly in this instance, 'the problem [would become] that of assessing which result is more accurate on the basis of the different merits of each model' (Lisciandra 2017, 88). And if this is so, I don't see why the nature of the problem should change when it comes to Weisberg and Reisman's individual based model's result. In other words, it seems prima-facie reasonable to assume that the epistemic value of learning that an individual-based model also manifests

[13] Notice that since the models do not share a mathematical framework it is hard to see how Raz (2016)'s generalization approach could work in this case.

the Volterra property should be determined on the basis of the merits of this model, and this model alone.

Overall, from the above discussion it is clear that the popular claim that robustness analysis allows one to learn that a model's result is not an artefact of a particular idealization/assumption is a particularly ambiguous one. In particular, in this section we have seen that such a claim *must* be interpreted differently depending on what particular instance of robustness analysis one is dealing with in any given case. And we have further seen that the claim must in fact be false as far as some instances of robustness analysis are concerned (e.g. cases of representational robustness analysis). But if we are genuinely interested in understanding the epistemic import of robustness analysis, it is important that we desist from relying on ambiguous claims in our efforts to do so. Hence in what follows, I shall endeavour to avoid any such ambiguous claims.

3.3 Robust theorems, low-level confirmation and ceteris paribus clauses

According to Weisberg (2006) the discovery of the Volterra principle through the analysis of predator-prey models provides an 'excellent template for a more general characterization of robustness analysis' (ibid. 737), which he characterizes as a four step procedure: (i) evaluate whether a group of models share a common result R; (ii) determine whether this set of models share a common substantial assumption C; (iii) formulate the robust theorem: a conditional statement linking the common substantial assumption C to the robust property R, prefaced by a ceteris paribus clause; (iv) conduct "stability analysis" of the robust theorem, with the aim of finding out what conditions will defeat the connection between C and R.

In light of Weisberg's general characterization of robustness analysis, there are a couple of questions on which I want to focus in this section. First, what does a robust theorem say about the world (i.e. what is a robust theorem's empirical content) according to Weisberg? Second, what reasons do we have for believing that the robust theorem is a *true* theorem about the world according to

Weisberg? Before grappling with these questions, however, let us have a closer look at what Weisberg has to say about each of the above steps.

3.3.1 Weisberg's general characterization of robustness analysis

As mentioned above, in the first step, one must evaluate whether a set of models share a common result *R*. All that Weisberg says about this step is to make sure to collect 'a sufficiently diverse set of models so that the discovery of a robust property does not depend in an arbitrary way on the set of models analyzed' (ibid. 737). Weisberg, however, does not clarify what it takes for a set of models to be *sufficiently* diverse for this step to be carried out appropriately, so if there are any criteria on which to select these models according Weisberg, these are at best left vague. In the second step (conducted subsequently to the first step or in parallel with it) one must determine whether this set of models share a common structure. However, Weisberg notes that this step might not always be straightforward to carry out. For although there are cases in which the common structure will have the very same mathematical structure in each model (as, for instance, we have seen in the case of parameter and structural robustness for the Volterra principle), this may not always be the case. For when models are developed using different mathematical frameworks, or 'represent a similar causal structure in different ways or different levels of abstraction' (ibid., 738), if the models are deemed to share a common structure this cannot be due to them sharing the very same mathematical structure. Hence:

> Such cases are much harder to describe in general, relying as they do on the theorist's ability to judge relevantly similar structures. In the most rigorous cases, theorists can demonstrate that each token of the common structure gives rise to the robust behavior and that the tokens of the common structure contain important mathematical similarities, not just intuitive qualitative similarities. However, there are occasions in which theorists rely on judgment and experience, not mathematics or simulation, to make such determinations. (ibid., 738)

If the first and second step have been carried out successfully, one can proceed to the third step: formulating the robust theorem of the general form: 'Ceteris paribus, if [common causal structure] obtains, then [robust property] will obtain' (ibid., 738). Weisberg stresses that in order to carry out this step successfully one must interpret the common structure shared by the models and the robust property as descriptions of empirical phenomena. For if the robust theorem is indeed a theorem about the real world, then it must concern properties of real-world phenomena, not of mathematical structures.

In the fourth and final step, one should conduct 'various kinds of stability analaysis'. The purpose of this step according to Weisberg 'is to determine what happens to the robust theorem when the situation described by the set of models varies slightly' (ibid., 738). As an instance of stability analysis, Weisberg asks us to the consider the transition from the original Lotka-Volterra model and the one with the prey population growth rate density dependent (both models were discussed in Section 3.3.1). One way to think about this transition is to ask whether the Volterra principle will still hold, 'when density dependence, even an arbitrary small amount of it, is factored into the model' (ibid., 738). As discussed in Section 3.3.1, the Volterra principle turns out to be insensitive to density dependence (because it holds for all parameter values). According to Weisberg, in cases where stability analysis is carried out extensively, 'it may ultimately be possible to replace a robust theorem's general ceteris paribus clause with a very specific statement of the conditions that defeat the efficacy of the core structure in generating the robust properties' (ibid. 739).

So in a nutshell, robustness analysis, according to Weisberg is a four step procedure that allows as to discover robust theorems of the general form: 'Ceteris paribus, if [common causal structure] obtains, then [robust property] will obtain', which are theorems about the real world. But what can we do with these theorems? Does Weisberg think we can use them to explain a real-world phenomenon or to predict its occurrence? Yes and no, according to Weisberg. No, because robust theorems are conditional statements, further attenuated with ceteris paribus clauses. But for robust theorems to allows us to give us an adequate

explanation of a real world phenomenon or a successful prediction of its occurrence, we would have to know that the common structure is actually being instantiated in the target system and that no other causal factor is preempting the efficacy of the common structure. Since robust theorems are silent with respect to these questions, they can't on their own increase the quality of our predictions and explanations about real-world phenomena. Yes, because if were we to know that the common structure is instantiated and that no preempting causal factors are present in the target system, we could in such cases use robust theorems to explain or predict.

Weisberg seems to accept that whether the common structure is being instantiated and if any preempting causes are present in the target system can only be reliably assessed though an empirical investigation. However, he also suggests that in cases where it is impossible to collect the relevant data, there are some techniques associated with robustness analysis that can help us settle whether the common structure is instantiated in the target system and that no other causal factor that can preempt its efficacy is present.

How can techniques associated with robustness analysis help us settle that no other causal factor is preempting the efficacy of the common structure? According to Weisberg the answer lies in the fourth step of robustness analysis:

> In order to determine how sensitive a robust property is to perturbations, theorists engage in various kinds of stability analyses. If fully carried out, the fourth step of robustness analysis provides enough information to determine what kinds of perturbations will preempt the occurrence of the robust property, even when the core structure is instantiated. (ibid., 740)

However, notice that even if, through stability analysis, we were to achieve an understanding of *all* the conditions that defeat the efficacy of the core structure in generating the robust property, without the knowledge of whether or not those conditions are present in the target system, this is not going to help us settle that no causal factor is preempting the efficacy of the common structure in the target system. In order to settle this, it seems to me, we would also have to know

that those conditions are not present in the target systems, and this can only be settled though an empirical investigation of the real phenomenon of interest.

How can techniques associated with robustness analysis help us settle whether or not the common causal structure is instantiated in the target system? According to Weisberg:

> The key comes in ensuring that a sufficiently heterogeneous set of situations is covered in the set of models subjected to robustness analysis. If a sufficiently heterogeneous set of models for a phenomenon all have the common structure, then it is very likely that the real-world phenomenon has a corresponding causal structure. This would allow us to infer that when we observe the robust property in a real system, then it is likely that the core structure is present and that it is giving rise to the property. (ibid., 739)

Weisberg makes two claims in the above quote. The first is that 'if a sufficiently heterogeneous set of models for a phenomenon all have the common structure, then it is very likely that the real-world phenomenon has a corresponding causal structure'. The only way to make sense of this claim, it seems to me, is if according to Weisberg a sufficiently heterogeneous set of models is one which exhaustively (or nearly exhaustively) samples all possible representations of the target system. If this were the case, then the discovery that all these models in this set happen[14] to share a causal structure would seem to entitle us to infer that that causal structure must be instantiated in the target system. Weisberg's second claim is that 'this would allow us to infer that when we observe the robust property in a real system, then it is likely that the core structure is present and that it is giving rise to the property'. Weisberg's idea here must be that because this set of models exhaustively samples all possible representations of the target system, and we have identified only one causal structure that can give rise to

[14]I use the term 'happen' here to stress the fact that the models in this set couldn't have been selected on the basis that they share a common causal structure. For if they were selected on such basis, then it is plausible to believe this set of models exhaustively samples all possible representations of the target system only if one has independent reasons for believing that any possible representation of the target system must necessarily have that causal structure. But if this were the case, one would already believe that the causal structure is present prior to subjecting any model to robustness analysis!

the robust property across all these models, then if we observe the robust property in a real system, we seem to be entitled to infer that the observed property must be due to that causal structure being present and giving rise to it (since we have determined that no other causal structure can give rise to that property)[15]. In other words, 'The qualifier "sufficiently heterogeneous" helps guard against the possibility that another structure found outside the set of models considered generates [the robust property]' (Justus 2012, 800).

I suggested that the only way to make sense of Weisberg's two claims above is that a sufficiently heterogeneous set is one that exhaustively (or nearly exhaustively) samples all possible representations of the phenomenon of interest for as Houkes and Vaesen (2012, 352) put it 'that instantiation of the causal structure is only credible if, as the formalist critics submit, all conceivable models of a target system or phenomenon are inspected. Until this completeness has been achieved, any shared structure found responsible for a robust property may be an artefact of the limited scope of explorative robustness analysis, even if the implications of highly diverse models would be inspected'. But if this is what it takes for a set of models to be sufficiently heterogeneous, then Weisberg's notion of sufficient heterogeneity is clearly an extremely demanding one, one that is arguably very hard (if even possible) to achieve in most cases. But without achieving it, the inference from the observation of the robust property in a real system, to the hypothesis that the common causal structure is instantiated and that it is giving rise to that property would *at best* be an inference to the best explanation (or rather an inference to the only possible explanation that we have discovered so far). It is also important to recognize that this notion of 'sufficient heterogeneity' is in fact irrelevant to the practice of robustness analysis as characterized by Weisberg. This is because one of the conditions for the set of models to be subjected to robustness analysis is that the models share a common structure. In other words, if we were to find out that a model does not have the common structure shared by the other models we would regard the behavior of this

[15]Under this interpretation, however, it is unclear why the second claim relies on the first claim. For the second claim would still hold if not all models for a phenomenon were to share the same causal structure, as long as the models that do not have the causal structure in question do not also have the robust property.

model as irrelevant for our current purpose, that of discovering a robust theorem (which concerns a given causal structure and a given property that this structure is supposed to give rise to). For instance, the only models that are relevant to assess the robustness of the Volterra principle are models of predator-prey systems that are negatively coupled, because 'negative coupling is a necessary condition for a system to demonstrate the Volterra Principle' (Weisberg and Reisman 2008, 124).

Overall, despite Weisberg's (tentative) suggestions on how techniques associated with robustness analysis can help us settle whether a causal structure is instantiated in the target system and whether any causal factor is preempting the efficacy of the common structure, I will assume that these are questions that should be settled through an empirical investigation and that the practice of robustness analysis is not about providing an answer to these questions. Given this, it is time to come back to the two questions that I raised at the beginning of this section:

1. What does a 'robust theorem' say about the world (i.e. what is the empirical content of a robust theorem)?

2. Why should we believe a robust theorem is a *true* claim about the world?

In the next section, I will argue that by accepting Weisberg's answer to the second question, it is impossible to give an adequate answer to the first; and that it is unclear what is the epistemic value of having an answer to the second question, without having an answer to the first.

3.3.2 On the empirical content of robust theorems

What does Weisberg have to say about the second question? That is why, according to Weisberg, should we believe a robust theorem is a *true* claim about the world?

According to Weisberg the robust theorems that are generated in the the third step of robustness analysis are *confirmed* empirical hypotheses. But what entitles us to move from a mathematical fact (i.e the fact that a set of models that share a common structure all entail a result) to an empirical one? Weisberg argues

that although the move may appear illicit for it seems to rely on some sort of "nonempirical confirmation", it is not:

> While the transition from mathematical to empirical may look illicit when described as "nonempirical confirmation," it is actually part of a well-accepted theoretical practice that is so common, it is rarely discussed explicitly. In every scientific domain, theorists must establish that the mathematical framework in which their theories are framed can adequately represent the phenomena of interest. (ibid., 740)

According to Weisberg, what licenses us to move from the mathematical to the empirical is what he calls "low-level confirmation", which is 'the sort of confirmation that licenses the use of framework to construct models of phenomena in the first place' (ibid., 742). As an example of low-level confirmation, he considers the logistic growth model of population. According to Weisberg, if we know that a population is growing logistically, the very fact that we think that the logistic growth model adequately represents this growth relies on low-level confirmation:

> By way of example, consider models of population growth. Standard issues in confirmation theory concern whether a particular kind of model, such as the logistic growth model, is confirmed by the available data. However, there is a prior confirmation-theoretic question that is often asked only implicitly: If the population is growing logistically, can the mathematics of the logistic growth model adequately represent this growth? Theorists rarely articulate such questions in research articles, but an affirmative answer underlies their research. (ibid., 740)

Similarly, Weisberg argues that what confirms the robust theorems discovered through the practice of robustness analysis is the low-level confirmation of the mathematical framework in which they are embedded:

> In the predation case, for example, we are confident that ecological relationships can be represented with the models described by

> coupled differential equations. Thus when we discover the consequences of these models, we are confident that most of these consequences are true of any system described by the model[s]. This confidence comes from low-level confirmation, not from robustness analysis itself. Thus robustness analysis is not a nonempirical form of confirmation as Orzack and Sober suggest. It does not confirm robust theorems; it identifies hypotheses whose confirmation derives from the low-level confirmation of the mathematical framework in which they are embedded. (ibid., 741)

However, Weisberg's notion of low-level confirmation does raise some questions and some hesitations too. For a start, one may worry that Weisberg's notion of low-level confirmation relies on the idea that mathematics itself is confirmable (as Justus (2012) does). If so, this would be odd for more than one reason:

> Apart from the unusual idea that mathematics itself is confirmable (Sober 1993), however low level, it is unclear whether this modality can do that work. First, it seems highly implausible that richer, more expressive mathematical frameworks are somehow confirmed by their greater representational capability. What matters for confirmation is whether an empirically interpreted mathematical structure does adequately represent, not whether the mathematical framework in which it is expressed can. Second, confirmation increases probability for almost all theories of confirmation. If mathematical frameworks possess low-level confirmation, this would entail the odd result that empirically interpreted but thoroughly empirically inadequate mathematical structures would nevertheless receive a probability boost from the representationally adept framework they are expressed within'. (Justus 2012, 800)

Luckily, however, if all that low-level confirmation is supposed to do is allow us to give an affirmative answer to prior confirmation-theoretic questions such as 'if the population is growing logistically, can the mathematics of the logistic growth model adequately represent this growth?', then this does not seem to

rely on the idea that maths itself is confirmable (whatever that may mean). So, arguably, this worry is unjustified.

Notwithstanding this, there is a second more important worry. As Houkes and Vaesen observe, Weisberg is ambiguous about the scope in his notion of low-level confirmation. Is it supposed to apply to a broad mathematical framework, say that of coupled differential equations? Or to a specific model family? Or to individual models?:

> Weisberg (2006, 740–41) suggests that low-level confirmation is based on predictive accuracy and that it warrants belief in the representational accuracy of both "the mathematics of the logistic model" and "the models described by coupled differential equations." This illustrates, in our opinion, the ambiguities of scope in the notion of low-level confirmation: even if one assumes that the former applies to all models described by the logistic equation, it is much more specific than the latter—and one would expect such differences to be relevant to the scope of robustness analysis. (Houkes and Vaesen 2012, 353)

Third, and relatedly, if low-level confirmation is the sort of confirmation that licenses the use of framework to construct models of phenomena in the first, then what is the feature/property of the framework that we compare to reality to determine when we are indeed licensed to do so? Without a clear understanding of the scope in the notion of low-level confirmation, it seems particularly hard to give an adequate answer to this question.

Last but not least, it is important to note that according to Weisberg low-level confirmation licenses us to believe that for all the models we have collected to undergo robustness analysis 'when we discover the consequences of these models, we are confident that most of these consequences are true of *any system described by the model[s]*' (Weisberg 2006, 741; my emphasis). This means that if the systems described the models are fictional systems, that is systems that are unrealistic with respect to the target system in various respects, then all that low-level confirmation allows us to establish, according to Weisberg, is that if those

fictional systems happened to exist in the real-world then we would be entitled to believe that the consequences of our models are true in those systems. But then, if we are interested in learning about properties of the target system and not a fictional one, low-level confirmation in and of itself can't help us with that.

In any case, and independently of what low-level confirmation is all about, if as Weisberg argues low-confirmation is really what entitles us to believe the robust theorems to be true theorems about the world, and robustness analysis is merely a procedure to discover them, a question arises: why does Weisberg stress that one should collect 'a sufficiently diverse set of models so that the discovery of a robust property does not depend in an arbitrary way on the set of models analyzed' in the first step of robustness analysis? In other words, if low-level confirmation is really what 'licenses us to regard the mathematical dependence of [property R] on [structure C] as a causal dependence' (ibid., 741) then it seems that whether or not this mathematical dependence should be regarded as a causal dependence should not depend on the diversity of the set of models we collect in the first step of robustness analysis. If it did then something *other than* low-level confirmation would have to be playing a role in the confirmation of robust theorems. This would be in conflict with Weisberg's own justification for why we should believe robust theorems.

So why *do* we need a sufficiently diverse set of models in the first step of robustness analysis? This brings us to the first question I raised above: what is the empirical content of robust theorems? Even on the assumption that low-level confirmation licenses us to believe the robust theorems to be *true* empirical theorems, as suggested by Weisberg, there is the further question as to what is the empirical content of a robust theorem in the first place. Recall that the robust theorems generated in the third step of robustness analysis are hypotheses of the general form: 'Ceteris paribus, if [common causal structure] obtains, then [robust property] will obtain.' But how should we interpret these hypotheses given the role of the ceteris paribus clause in them?

The question of the determination of a clear interpretation of 'ceteris paribus' (cp) clauses has received considerable attention in the philosophical literature. Part of the motivation behind philosophers' interest in this question has to do

with the concern that cp laws appear to lack empirically testable content, that is they appear to be analytically true sentences (and hence trivially true) rather than empirical statements. The problem concerning cp laws is usually posed as a dilemma (originally formulated by Lange (1993)): if cp laws are reconstructed as strict generalizations then they are bound to be false (since just one counter-instance is needed for them to be false, and typically one is not hard to find); on the other hand if we assume that cp laws are not strict generalizations but claims of the form "all As are Bs, if nothing interferes" then they seem to say nothing more than "(all As are Bs) or not (all As are Bs)" which are analytically true statements devoid of any empirical content.

Earman et al. (2002) make an important distinction between "lazy" and "non-lazy" cp-clauses. A lazy cp clause is one that is effectively dispensable because all the conditions that have to obtain for the generalization to be true are in fact known, but not listed explicitly merely as a result of "laziness". Hence lazy cp laws *can* avoid the horns of Lange's dilemma. A non-lazy cp clause, on the other hand, is not dispensable because a complete list of all the conditions that have to obtain for the generalization to be true is impossible (e.g. due to the list being infinite or open ended). And indeed, a complete description of all possible conditions that have to obtain for a generalization to be true is often impossible. Consider, for instance, the claim 'ceteris paribus, humans can swim'. There are an infinite number of factors that may affect a human's ability to swim. Hence, there will always be a counter instance to the claim that 'humans satisfying C can swim' for every condition C which excludes a finite list of such factors. Hence a strict completion of all possible conditions that have to obtain for this generalization to be true is impossible. This is one of the many instances of a "non-lazy" cp clause and I think there are good reasons to think that the cp clause in the robust theorems generated by the practice of robustness analysis is also non-lazy (i.e. indispensable). But if this is right, how to dismiss the worry that those theorems are devoid of any empirical content?

Clearly, Weisberg does not intend robust theorems to be analytically true statements. But in order for this to be the case we must find a way to interpret the cp clause in such a way that it does not render robust theorems trivially

true. One promising attempt is offered by Lange (2000) (see Reutlinger et al. (2021) for a review of various other attempts).[16] The essential idea in Lange's attempt to rescue non-lazy cp laws is to restrict their application to the purposes of a scientific discipline. In particular, according to Lange the cp clause should be treated as a name for a set all intervening factors I that are *relevant* (for a particular discipline) and only those. These factors are relevant if 'they arise sufficiently often, and can cause sufficiently great deviations from G-hood, that a policy of inferring Fs to be G [...] would not be good enough for the relevant purposes' (Lange 2000, 170) and fall into the range of the laws intended purpose and application.

Hence by treating the ceteris paribus clause as a set I of all intervening factors that are relevant and only those, we can perhaps rescue robust theorems from analytic triviality. However, notice that without an understanding of the situations described by the set of models collected in the first step of robustness analysis, we would be forced to treat the 'ceteris paribus' clause as a name for the set of *all* the factors that arise sufficiently often and that fall in the range of the law intended purpose. For if we don't know whether those factors will disrupt the efficacy of the robust theorem then we are not entitled to assume that they won't. This move, however, would essentially render robust theorems only empirically informative about worlds that we have extremely good reasons to believe we are not in. Suppose for instance that we didn't know that the Volterra principle is insensitive to density dependence. In this case we could not assume that the Volterra principle concerns any system with density dependence, no matter how small. But a Volterra principle which concerns only predator-prey systems with no density dependence at all, is arguably not a theorem about the actual world, because any real system is bound to have *some* density dependence. As another example (not considered by Weisberg and Reisman), if we don't know whether or not the Volterra principle is sensitive to various predators and prey's

[16] Another major attempt to rescue cp laws from analytic triviality is to claim that these laws are meant to reveal dispositions, and dispositions can be instantiated without being manifested (Cartwright 1989, Lipton 1999, Hüttemann 2014). Hence, so the thought goes, when cp laws are understood as laws that ascribe dispositions, rather than regular behaviour, they are strict true laws and can avoid the horns of Lange's dilemma. As I will discuss later on in this section, this contrasting attempt to rescue cp laws might help us provide an alternative interpretation of robust theorems.

responses to seasonal fluctuations, we cannot assume that the Volterra principle concerns any predator-prey system in a seasonal environment. And yet virtually all real-world biological populations live in a seasonal environment. Arguably, this might be the very worry that underlies Weisberg's recommendation to collect 'a sufficiently diverse set of models so that the discovery of a robust property does not depend in an arbitrary way on the set of models analyzed' in the first step of robustness analysis. But this is not helpful to dismiss this worry. For no matter how diverse the set may be, only an understanding of the situations that are covered by the set of models can enable us to remove from the set I some of those factors that arise sufficiently often and that are relevant for the theorem intended purpose. Although the fourth step of robustness analysis (i.e. stability analysis) may (or may not) eventually help us remove some or even many of those factors from the set *I* (and hence increase the empirical informativeness of the robust theorem), the very fact that we can, according to Weisberg, declare to have a robust theorem prior to this step, must mean that robust theorems, under Weisberg's general characterization of robustness analysis, don't have to be theorems concerning the actual world to deserve the name.

This matters. For if robust theorems do not have to concern the actual world for them to deserve the name, then regardless of whether or not those theorems are confirmed (by low-level confirmation), it cannot be assumed they will *ever* be useful for explaining or predicting real-world phenomena. For there is nothing in Weisberg's notion of robust theorems that licenses us to assume that they *can* concern the actual world. Hence, it seems to me that by accepting that low-level confirmation automatically confirms 'robust theorems' as Weisberg argues one should, one would also have to accept that whether a robust theorem can be useful for explaining or predicting real-world phenomena is irrelevant to whether we choose to call it a robust theorem. Or to put it in other words, by accepting that low-level confirmation automatically confirms 'robust theorems', one would also have to accept that robust theorems *do not and might never* 'establish conduits through which empirical support for *C* can transmit to *R*, and vice versa' (Justus 2012, 800), contrary to what has often been assumed.

To recapitulate, according to Weisberg robustness analysis is a procedure

which can be used to discover robust theorems, which are empirical theorems of the general form 'ceteris paribus, if causal structure obtains robust property will obtain.' Robustness analysis, according to Weisberg, does not confirm those theorems. What confirms them is the low-level confirmation of the mathematical framework in which they are embedded. However, I have argued that by accepting that 'low-level confirmation' automatically confirms 'robust theorems', as suggested by Weisberg, one must at the same time accept that robust theorems do not have to concern the actual world for them to deserve the name, and hence, one is not warranted to assume they can be useful for explaining or predicting real-world phenomena, contrary to what is usually assumed in the literature on robustness analysis. Hence if one thinks that robust theorems *should* concern the actual world, one cannot assume that they are automatically confirmed by Weisberg's notion of low-level confirmation.

Before concluding, it is worth mentioning that some of Weisberg's own remarks do in fact suggest that he may have a rather different interpretation of robust theorems than the one I have argued is warranted by his assumption that robust theorems are automatically confirmed by his notion of low-level confirmation. Consider for instance why, according to Weisberg, robust theorems on their own are insufficient for explaining or predicting a real world-phenomenon:

> Explaining a real-world phenomenon or predicting its occurrence requires us to know that the common structure is actually being instantiated and that no other causal factor is preempting the efficacy of the common structure. (Weisberg 2006, 739)

But if according to Weisberg the only reason why a robust theorem is insufficient for explaining or predicting a phenomenon, when we know that the common causal structure is instantiated, is that there may be some other causal factors that are 'preempting the efficacy of the common structure', then this suggests that, according to him, the efficacy of the common structure is always present, despite the fact that it may be preempted by some other causal factors. In other words, Weisberg doesn't seem to consider the possibility that this efficacy may

also be lost altogether. But if this is right then it seems that, according to Weisberg, robust theorems are in fact supposed to be interpreted as claims about (stable) capacities,[17] which are introduced by Cartwright to explain causal laws and render them universal in character: if *C* has the (stable) capacity to produce *R* then *C* carries this capacity from situation to situation (Cartwright 1989, 145).

Under an interpretation of robust theorems as claims about (stable) capacities, a case could perhaps be made for why we should think of them as relevant to the explanation and prediction of real-world phenomena. For under this interpretation, robust theorems describe how *real-world* systems behave in the absence of disturbing factors and this knowledge can in principle be 'used to account for more complex situations, in which various systems and their dispositions are intertwined—provided laws of superposition are available' (Reutlinger et al. 2021, 32). In any case, my aim here is not to defend the view that the best interpretation of cp-laws or robust theorems is as claims about (stable) capacities. The sole aim of this discussion is to stress that although Weisberg himself may implicitly think of robust theorems as claims about (stable) capacities, which could perhaps shed light on why we should deem them to be relevant to the explanation and prediction of real-world phenomena, this interpretation of robust theorems is not compatible with Weisberg's claim that the role of robustness analysis is merely to discover robust theorems, theorems that are automatically confirmed by his notion of low-level confirmation. This is because, as I have argued above, robust theorems, under the view that they are automatically confirmed by low-level confirmation, do not have to concern the actual world

[17] If a cp law is interpreted as a law about capacities then it would seem that a cp law 'is no longer considered to be a description of the systems' occurrent behaviour that is only manifest under very special conditions—if at all. The law concerns the underlying stable tendencies or disposition' Reutlinger et al. (2021). So why am I adding a stable qualifier? As Schrenk (2007) convincingly argues, in sciences further down the hierarchy than fundamental physics, capacities or dispositions cannot only fail to be manifested, but can also 'be lost because their underlying basis breaks or alters. Haemoglobin cells might be damaged and not be able to bind O2 anymore, birds might lose their ability to fly because their wings are broken, [...] etc.' (ibid., 17). Hence, in these cases, laws about capacities would still need a ceteris paribus that stands for the presence or absence of the capacity. Indeed, Cartwright herself is also sceptical about the existence of *stable* capacities in the social realm in particular, since 'economic features have the capacities they do because of some underlying social, institutional, legal and psychological arrangements that give rise to them. So the strengths of economic capacities can be changed, unlike many in physics, because the underlying structures from which they derive can be altered' (Cartwright cited in Crespo 2013, 28). This is why I have added a stable qualifier to capture how Weisberg seems to implicitly interpret robust theorems.

(including facts about stable capacities) for them to deserve the name. In other words, if according to Weisberg robust theorems really are meant to be interpreted as claims about stable capacities, then the idea that his notion of low-level confirmation is able to automatically confirm those theorems is untenable.

As we will see in the next chapter, some philosophers have argued that the role of robustness analysis is not merely to discover robust theorems, but to confirm them too. This must mean that, in their view, the robust theorems can be confirmed even if/when Weisberg's notion of low-level confirmation cannot. Let us see whether they can convince us.

Chapter 4

Robustness analysis as a tool for confirming robust theorems: an assessment of some popular arguments

4.1 Introduction

The overall aim of this chapter is to critically assess the validity and soundness of various distinct arguments that have been offered to motivate the idea that robustness analysis (RA)s has a rightful claim as a method of confirmation of a 'robust theorem'. Its structure is as follows. In Section 4.2, I will critically assess an argument put forward by Kuorikoski et al. (2010) for the epistemic import of model-based RA, an argument which I believe is a formal expression of a widely held but ultimately misleading intuition: the intuition that a model's conclusion is more likely to hold in the target system if several models lead to that conclusion because it would be a remarkable coincidence if that were not the case. Kuorikoski et al. offer the best available defence of this intuition and that is why I believe it is important to rigorously assess it. I will conclude that, although Kuorikoski et al.'s argument relies on a weaker notion of probabilistic independence than unconditional probabilistic independence, it cannot be sound. By relying on a different notion of independence (i.e. Fitelson's (2001) account of confirmational independence), I will then offer a revised, prima-facie

more plausible argument. However, I will conclude that this argument also relies on assumptions that are hardly ever plausible. This strongly suggests that any successful argument to support the idea that RA is able to confirm a 'robust theorem' cannot rely on any sort of probabilistic independence to explicate the notion of model diversity. In Section 4.3, I will turn to Schupbach's (2018) recent account of RA as explanatory reasoning. I will show that, although this account seems to fit well with some empirical cases of RA, when one tries to apply Schupbach's account to model-based RA the picture appears rather more complicated than Schupbach suggests, as its application relies on several non-trivial assumptions. Despite this, I will argue that those assumptions may be reasonable in cases where the hypothesis we are interested in confirming through model-based RA is a 'robust theorem'. Hence, I will conclude that Schupbach's account could indeed be an adequate (Bayesian) account for justifying why and determining when model-based RA should increase one's confidence in a 'robust theorem', and also for helping us understand the extent of that confirmation.

4.2 The epistemic value of independent lies: false analogies and equivocations.

The aim of this section is to critically assess an argument put forward by Kuorikoski et al. (2010) for the epistemic import of model-based robustness analysis. This assessment is important for two reasons. First, I believe Kuorikoski et al.'s argument is a formal expression of a widely held, but what I believe to be an ultimately misleading, intuition. This intuition is the following: a model's conclusion is more likely to hold in the target system if several models lead to that conclusion because it would be a remarkable coincidence if that were not the case. Kuorikoski et al. offer the best available defence of this intuition and that is why I believe it is important to rigorously assess it. Second, several arguments for the epistemic import of robustness analysis that have been offered so far are neither formulated nor defended with sufficient clarity and precision. Hence, in

my view, a serious investigation into the epistemic import of robustness analysis must start with a careful reconstruction of those arguments, followed by a rigorous assessment of the tenability of the premises of those arguments. The purpose of this section is to critically assess Kuoriskoski et al.'s argument in particular; I will conclude that the assumptions on which this argument relies are implausible. I must point out that I am not the first to object to Kuorikoski et al.'s (2010) argument. Odenbaugh and Alexandranova (2011) have also questioned the validity of some of its assumptions. However, in my view, their objections were insufficient, and thus so were Kuorikoski et al.'s (2012) responses. Here, I hope to show more forcefully that the assumptions that underscore Kuorikoski et al.'s argument are untenable.

For the purpose of this discussion, I will assume that the 'substantial assumptions' in a model are those that 'identify a set of causal factors that in interaction make up the causal mechanism about which the modeller endeavours to make important claims' (Kuorikoski et al. 2010, 547). Following Kuorikoski et al., I will assume that there are two conceptually distinct kinds of idealizations: Galilean assumptions and tractability assumptions. Galilean assumptions 'serve to isolate the working of the core causal mechanism by idealising away the influence of the confounding factors' (ibid., 547). According to Kuorikoski et al., despite being unrealistic with respect to the model's target system, Galilean assumptions have a causal interpretation: 'they state that a factor known or presumed to have an effect is absent' (ibid., 547). Tractability assumptions, on the other hand, are assumptions that are introduced 'only for reasons of mathematical tractability' and, in contrast to Galilean assumptions, they often 'have no empirical merit on their own' (ibid., 548) and hence 'the falsehood they embody is hoped to be irrelevant for the model's result' (ibid., 548). This is why, according to Kuorikoski et al., 'unlike Galilean idealisations, for many tractability assumptions it is often unclear what it would mean to replace them with more realistic ones: if it were possible to do without these kind of assumptions they would not be introduced in the first place' (ibid., 548). Throughout this section, I will denote the substantial assumptions by C, all the Galilean assumptions by G and all the tractability assumptions by T.

As an illustration of this working definition, take the Lotka-Volterra model, discussed in Section 3.2. In line with discussions of this model in the literature on robustness, I will take the substantial assumption in this model to be the assumption that the target predator-prey system is negatively coupled (i.e. increasing the size of the predator population decreases the size of prey population and increasing the size of the prey population increases the size of the predator population). And in line with the definition given above, an example of a Galilean assumption could be the assumption that aside from the size of the predator population, there are no other factors that may affect the size of the prey population (such as limited resources). Notice that although this is an unrealistic assumption with respect any real-world predator-prey system, it could in principle be replaced with a more realistic assumption; for instance by replacing it with the assumption that there is a maximum carrying capacity to the growth rate of the prey population as done by Weisberg (2006) and discussed in Section 3.2. According to Kuorikoski et al. (2012), an example of a tractability assumption could be the *specific* functional form used to describe the rate of prey capture per predator (this model assumes that there is a linear increase in prey capture with prey density). Kuorikoski et al. (2012) consider this to be a tractability assumption in so far as *any* assumed functional form for the rate of prey capture will 'strictly speaking be false for any natural population' (ibid., 8).[1]

Kuorikoski et al. (2010) largely agree with Weisberg's characterization of robustness analysis. However, they argue that the failure of robustness with respect to tractability assumptions is epistemically problematic 'because it suggests that the result is an artefact of the specific set of tractability assumptions, which in many cases have no empirical merit on their own' (ibid., 548). What

[1]This last claim may strike the reader as being a little strong since it certainly seems possible, in principle, that a particular assumed functional form could be true. Crucially, however, even if any assumed functional form for the rate of prey capture is unlikely to be strictly true, there is certainly a sense in which one particular functional form could be more approximately accurate than another. And if this is the case, then it is not clear why one should think of these assumptions (i.e. specific choices of functional forms) as being introduced 'only for reasons of mathematical tractability', as Kuorikoski et al. seem to suggest. It is also worth pointing out that Kuorikoski et al.'s (2010) case study is not the Lotka-Volterra model, but a model in geographical economics. According to them, examples of tractability assumptions in that case are 'specific functional forms of utility [...], production [...] and transformation technology [...]' (ibid., 556). It seems to me that the above considerations should apply to these examples too. I shall return to the question of how we should interpret tractability assumptions at the end of this section.

this means is that, in contrast to Weisberg, when models involve tractability assumptions (as they often, if not always, do), Kuorikoski et al. don't think that we are licensed to believe the robust theorems discovered through the practice of robustness analysis. In contrast, in their view, the failure of robustness with respect to Galilean assumptions is *not* epistemically problematic because 'it [merely] suggests a new empirical hypothesis about a causally relevant feature in the modelled system' (ibid. 552). Or to put it in other words, according to Kuorikoski et al. Galilean assumptions can effectively be packed into the cp clause and therefore do not affect the validity of the robust theorem. This is why, as we will shortly see, Kuorikoski et al.'s argument for the epistemic import of robustness analysis focuses exclusively on models that involve different tractability assumptions, while keeping constant all Galilean assumptions.

Before I get to Kuorikoski et al.'s argument, I need to make a few clarifications. As discussed in the previous chapter, the Volterra principle is meant to be an empirical hypothesis. However, I have also argued that due to the use of the ceteris paribus clause it is not very clear how one should interpret this principle. For the purpose of this section, and in line with how philosophers have, in my view, often implicitly interpreted the Volterra principle and robust theorems more generally (e.g. Weisberg (2006), Kuorikoski et al. (2010)), I will assume the Volterra principle is a causal hypothesis; that is, according to the Volterra principle, a two-species predator-prey system that is negatively coupled has 'the efficacy' (Weisberg, 2006) to produce the Volterra property, despite the fact that this efficacy may be preempted by possible intervening causal factors (which may or may not be present in a given predator-prey system). In particular, I will assume that the Volterra principle is a claim about capacities, which are introduced by Cartwright to explain causal laws and render them universal in character: if C has the capacity to produce R then C carries this capacity from situation to situation (Cartwright 1989, 145).[2] Although my objections to Kuorikoski et al.'s

[2]As mentioned in the previous chapter, Cartwright herself is sceptical about the existence of stable capacities in the social realm in particular, since 'economic features have the capacities they do because of some underlying social, institutional, legal and psychological arrangements that give rise to them. So the strengths of economic capacities can be changed, unlike many in physics, because the underlying structures from which they derive can be altered' (Cartwright cited in Crespo 2013, 28). Hence she worries that 'the license to move from the results in the model about what happens when a cause is exercised without impediment to a contribution that

argument will not ultimately rest on what particular interpretation of robust theorems one chooses, it is nonetheless important to stress that without a clear interpretation of the hypothesis we are trying to confirm, we clearly can't confirm it. Hence, my choice of interpretation of robust theorems, one that seems compatible with what Kuorikoski et. al.'s have in mind, should be seen as an attempt to clarify their argument for the epistemic import of robustness analysis and not as an attempt to restrict the scope of my objections.

I need to make one final clarification. If we care about explanation and prediction, we are clearly not merely interested in whether or not the Volterra principle is true. This is because even if interpreted as a claim about stable capacities, without knowing what causal factors can preempt those capacities from being manifested and if they are present in a particular prey-predator system, we cannot know whether those capacities can be manifested in that system. However, whether or not a causal factor may preempt a capacity from being manifested, although an important question for prediction and explanation, is an additional hypothesis that is independent of the truth of the Volterra principle. Hence, I will make the reasonable assumption that the question of whether or not some or many causal factors may preempt the efficacy of a negatively-coupled predator-prey system to produce the Volterra property is beyond the scope of Kuorikoski et al.'s argument for the epistemic import of RA in this case.

4.2.1 An argument from coincidence?

According to Kuorikoski et al. (2010, 560):

> Levins' (1966) unclear but intuitively appealing claim that 'our truth is the intersection of independent lies' could be taken to mean that result *R* can be derived from mechanism-description *C* using multiple

the cause will make in all situations of some designated category depends on the assumption that the cause has a stable contribution to make, and that assumption must be supported by evidence from elsewhere' (Cartwright 2009, 53). What I am assessing in this section, therefore, is whether Kuorikoski et al.'s account can show that model-based RA can provide some evidence for the assumption that a cause has a stable contribution to make.

> independent sets of untrue tractability assumptions. Various falsities involved in the different derivations do not matter if robustness analysis shows that result R does not depend on them.[3]

For Kuorikoski et al., the epistemic value of robustness analysis lies in the very *independence* of the different untrue tractability assumptions involved in the models, since if they are independent in the right sort of way, then (in their view) it can be shown that model-based robustness analysis is 'a species of general robustness analysis in the sense discussed by Wimsatt and that the same epistemic rationale applies to it' (ibid., 559). However, aside from mentioning that according to Wimsatt,

> [robustness] provides epistemic support via triangulation: a result is more likely to be real or reliable if a number of different and mutually *independent* routes lead to the same conclusion. *It would be a remarkable coincidence if separate and independent forms of determination yielded the same conclusion if the conclusion did not correspond to something real* (ibid., 544, my emphasis),

they neither clarify *what* is the epistemic rationale on which Wimsatt relies in his defence of the epistemic value of general robustness analysis, nor (as we will see in the next section) do they rely on it for their own defence of the epistemic value of robustness analysis. The sole aim of this section is to reflect on what to make of the very last line of the quote above: that it would be a remarkable coincidence if separate and independent forms of determination yielded the same conclusion if the conclusion did not correspond to something real.

Indeed, it is not hard to find cases where it would be a remarkable coincidence if the same conclusion of distinct forms of determination did not correspond to something real. Suppose, for instance, that I weigh myself on several distinct scales from different manufacturers and different suppliers and they all show that I weigh 300 pounds, a lot more than I thought I would. Despite this, I think to myself 'it would be too remarkable a coincidence if all these scales

[3] Kuorikoski et al. (2010) use the notation R_M to refer to a model's result, but to be consistent with my notation I replaced all instances of R_M with R.

showed that I weigh 300 pounds if I didn't really weigh 300 pounds. I must weigh 300 pounds!' No one should accuse me of irrationality here. But what kind of coincidence would this be? It would be the following: although each scale may mislead me, due to the possible presence of a faulty mechanism, I have no reason to suppose that these scales share the *same* faulty mechanism. Hence the fact that all these scales would mislead me in the same way for different reasons seems an extremely implausible concurrence of events. On the other hand, if my weight really was 300 pounds, and hence the scales' readings corresponded to something real (i.e. my weight), this concurrence of events would no longer seem a remarkable coincidence: under this hypothesis, all my scales are working well, and so through the right sort of causal mechanism my weight is causing the scales' readings to agree. Hence, it seems rational for me to opt for the hypothesis that does not involve a remarkable coincidence.[4]

Can one apply the same argument from coincidence that I applied to my scale example to the context of model-based robustness analysis? For this to be the case, one should be able to claim in this case too that it would be a remarkable coincidence if the same conclusion is implied by multiple models, each containing different tractability assumptions, if the conclusion did not correspond to something real *and* that the coincidence would vanish if the conclusion *did* correspond to something real. However, this is not the case. For, without further justification, the fact that these models all imply the same conclusion, *despite* each and every one of them containing false tractability assumptions, should still strike one as being a remarkable concurrence of events *even if* that conclusion were to correspond to something real. In other words, the fact that distinct models involving different false tractability assumptions give the same conclusion *is a coincidence*, but not one that seems to be explained away by the hypothesis that the conclusion corresponds to something real.

The crucial difference between my scale example and model-based robustness analysis is the following. In my scale example, we are able to postulate a

[4]Notice that this argument from coincidence crucially relies on the assumption that there is no *systematic error*, which seems reasonable in this case because all the scales come from different manufacturers and different suppliers. However, without this assumption, the convergence of the scales' readings would at best only entitle me to infer that that this convergence is not due to chance, but it would 'not indicate it is due to any specific cause.' (Mayo 1986, 45)

process that links the cause (i.e. my weight) to the effect (i.e. the scale's readings) and it is the very postulation of this causal process that explains why the scales' readings are the same. But in the case of models, we cannot postulate a causal process that links the reality of the conclusion to the models' conclusions. Scales are measuring instruments, they *measure* things through a *causal* process. Models are *not* measuring instruments, they don't measure things through a causal process; hence postulating that a model's conclusion is real is not enough to explain why distinct models agree on that conclusion. So it seems to me that, in order for the reality of the models' conclusion to help us explain away this coincidence, we would also have to tell a story about why the models that we are considering in a given case must *all* agree on that conclusion if the conclusion were to correspond to something real.[5] But whether that story can in fact be told does not seem to be something that can just be assumed. Indeed, consider a case where two models make incompatible assumptions about a specific target system. What story can one tell to justify the assumption that those two models would *have* to agree on a conclusion if that conclusion were to hold in that target system?

Perhaps, one could attempt to explain away this coincidence by merely appealing to the world of models and not the one outside them. But what would it mean to find a (non-causal) explanation for this coincidence in the world of models? One may be tempted to answer this question by simply noting that 'all models share a common core, which could be the main driver of the common conclusion' (an answer that I have more than once heard!). However, this assertion must be equivalent to the claim that a particular set of models which all share a common core all give the same conclusion. Now, if that set of models is the same set whose conclusion we have just observed, then this would be a tautological explanation: the explanation for why all the models in our ensemble give the same conclusion is that they all give the same conclusion. So this can't be right. If the claim is meant to appeal to a more general class of models of which our ensemble is but a subset, then this raises at least two questions:

[5] In my assessment of Schupbach's account of in the context of model-based robustness analysis (Section 4.2 and Section 5.3), I will discuss in more detail under what conditions such a story may or may not be plausible.

what is the relevant class of models? And in what sense would the fact that a more inclusive set of models all entail a conclusion provide an explanation for why a subset of it provides that conclusion? Alternatively, one might attempt to explain this coincidence by showing that the models in our ensemble are special cases of a more general model which gives the same conclusion. For instance, as discussed in the previous section, Raz (2017) demonstrates that as long as a condition that ensures that the average abundance of a system coincides with the relevant equilibrium is satisfied, the Volterra principle holds for a more general model. However, there are of course many cases where it cannot be shown that different models are special cases of a more general model, especially when models involve different representational frameworks (e.g. Weisberg and Reisman (2008) also consider an individual based model version of the Lotka-Volterra model). In any case, I think it is important to note that whether or not it is possible to find an explanation for this coincidence in the world of models, this explanation *alone* would not help us infer anything about the world outside of them (which is ultimately what we are interested in).

So there seems to be a prima-facie clear difference between my scale example and the example involving models: in the former a causal argument from coincidence for the truth of the conclusion seems to be justified, whereas the same cannot be said of the latter. Although Kuorikoski et al. do not advocate a causal argument from coincidence to defend their view about the epistemic import of robustness analysis (as we will see in the next section), they nonetheless do make several equivocatory remarks that nudge the reader in that direction. Consider, for instance, this passage:

> Before conducting robustness analysis we do not know for sure which part of the models is responsible for the result, although modellers usually have strong intuitions about this issue. If a result is implied by multiple models, each containing different sets of tractability assumptions, we may be more confident that the result depends not on the falsities we have introduced into the modelling, but rather on the common components [...]. Robustness analysis thus increases our

> confidence in the claim that the modelling result follows from the substantial assumptions, i.e. that some phenomenon can be caused by the core mechanism. (ibid., 551)

In the above quote, Kuorikoski et al. are suggesting that if a result is implied by multiple models, each containing different sets of tractability assumptions, our confidence that the result R *depends* on the common components (i.e. the substantial assumptions C), rather than the various different false tractability assumptions, should increase. At first glance, this reasoning may seem analogous to the reasoning that I applied to my scale example (i.e. a causal argument from coincidence). However on closer inspection, it is clearly based on an equivocation: one that, like most equivocations, has the potential to mislead. To see clearly why this is, it will be helpful to reconstruct Kuorikoski et al.'s above reasoning into a set of premises and a conclusion from those premises. Let M_i stand for a given model; the premises of Kuorikoski et al.'s argument are then the following:

$\boldsymbol{P_1}$: M_1 implies result R

$\vdots$

$\boldsymbol{P_n}$: M_n implies result R

By assumption, a model consists of substantial assumption C, Galilean assumptions G and tractability assumptions T. And, also by assumption, we are focusing on a class of models that all have the the same substantial assumptions and Galilean assumptions but that differ in their tractability assumptions. So the above premises can be rewritten as:

$\boldsymbol{P_1}$: $C\&G\&T_1$ implies result R

$\vdots$

$\boldsymbol{P_n}$: $C\&G\&T_n$ implies result R

According to Kuorikoski et al.'s above reasoning, from $P_1 \ldots P_n$, we are entitled to have more confidence in the following conclusion:[6]

[6]More confidence than the one we would have if we only had P_1.

Robustness conclusion (**R-C**): R depends on C.

In light of this argument, three observations are in place. First, notice that **R-C** is ambiguous between

R-C-model: In model land, R depends on C, and

R-C-world: In the actual world, R depends on C.

Second, given that all parts of a model are used in the derivation of a model's result, the only possible interpretation of **R-C-model** must be the following:

R-C-model: All models involving C in the relevant class imply result R.

But this interpretation of **R-C-model** is unclear without a specification of what is the relevant class of models. Are $M_1 \ldots M_n$ considered to be merely samples of this class or should we think of them as constituting the entire class? If the former, what is the relevant class? If the latter, why is this an interesting class? That is, why should we care about $M_1 \ldots M_n$? Without an answer to these questions it is really not clear how one should in fact interpret **R-C model**. As a side note, recall that according to Weisberg (2006, 739), 'if a *sufficiently heterogeneous* set of models for a phenomenon all have the common structure, then it is very likely that the real-world phenomenon has a corresponding causal structure' (my emphasis; Levins (1993) makes similar remarks). One might think that Weisberg's notion of sufficient heterogeneity is relevant to the questions I have just raised. However, and leaving aside the lack of clarity surrounding Weisberg's notion of sufficient heterogeneity, it seems to me that it is not in fact pertinent here. This is because, according to Weisberg, the purpose of robustness analysis is merely to 'identif[y] hypotheses' (ibid., 741) not to confirm them. Hence for Weisberg the only source of worry when it comes to evaluating the epistemic import of robustness analysis is the fact that the theorems generated by robustness analysis are 'conditional statements, further attenuated with ceteris paribus clauses' (ibid., 739). That is, the worry is that the robust theorem and its predictions hold only under certain conditions, but not in others. Hence the reason why we want a sufficiently heterogeneous set of models, according to Weisberg, is to address *this* worry: by considering models that satisfy various different conditions

we can raise our confidence that the theorem holds more generally. However, Kuorikoski et al.'s concern is of an altogether different nature. Kuorikoski et al. worry that due to the presence of tractability assumptions (which are assumed to be strictly false for any target system) the ceteris paribus theorem might not be a theorem about the real world in the first place. This is not to say that there is no answer to the questions I raise above, but it is to say that Weisberg's appeal to the notion of a sufficiently heterogeneous set of models should not be seen as an attempt to answer *those* questions. And we will see that Kuorikoski et al.'s argument for why robustness analysis should increase one's confidence in the ceteris paribus theorem also averts these questions altogether, by relying on a concept of independence instead.

Third, even if Kuorikoski et al. could give a clear interpretation of **R-C-model**, the transition from **R-C-model** to **R-C-world** needs to be justified. Doing this silently (as done in this argument) is a petitio principii because what needs to be shown is precisely that the transition from model land to the actual world is legitimate.

In the next section I will turn to what I consider to be Kuorikoski et al.'s official argument for the epistemic import of model-based robustness analysis. Indeed their 'official argument' could be interpreted as indirectly addressing these criticisms, so I will now turn to it.

4.2.2 What is Kuorikoski et al.'s argument?

Here is what Kuorikoski et al. write:

> Modelling can be considered as an act of inference from a set of substantial assumptions to a conclusion [...]. Tractability assumptions are typically needed for the process of inference to be feasible, but these assumptions may induce errors in the modelling process: they may lead us to believe falsities about the world even if the substantial assumptions are true. [...] We thus propose that the modeller should have no positive reason to believe that if one tractability assumption induces a certain kind of error (due to its falsehood) in the

result, so does another one. *Given that the modelling result of interest (R) is correct, prior probabilities concerning whether R can be derived from $C\&T_1$ or $C\&T_2 \ldots C\&T_n$ should be (roughly) independent. If the probabilities are independent in this way, then observing that the models lead to the same result rationally increases our degree of belief in the result*.[7,8] (ibid., 561 my emphasis)

There is a lot going on in this quote and I will need to introduce some new notation in order to unpack it. Let R_T be the proposition that result R is instantiated in the target system. And let R_k be the proposition that result R is derived by the kth model. From the above passage, the argument of Kuorikoski et al. for the epistemic import of model based robustness analysis seems to be the following:

The argument. Assume that we observe that a model with substantial assumption C and tractability assumptions T_1 gives result R. Then we will have some degrees of belief that the hypothesis h: "in the actual world, R causally depends on C" is true. Suppose further that in addition to our first model, we observe that several other models sharing the same substantial assumptions C, but differing in their tractability assumptions T_i give the same result R. This should rationally increase our degrees of belief in the hypothesis h, because it is reasonable to assume that the models' results R are probabilistically independent conditional on R_T (and $\neg R_T$).[9](i.e. because it reasonable to assume that $Pr(R_1\& \ldots \&R_n|R_T) = Pr(R_1|R_T) \times \ldots \times Pr(R_n|R_T)$ and $Pr(R_1\& \ldots \&R_n|\neg R_T) = Pr(R_1|\neg R_T) \times \ldots \times Pr(R_n|\neg R_T)$).

[7]It is clear from Kuorikoski et al.'s (2010) general discussion that "the result" at the end of this quote is not meant to refer to the hypothesis that R holds in the target system, but to the hypothesis that in the actual world, R causally depends on C. For instance, robustness analysis is supposed to increase our degrees of belief that the Volterra principle is correct, not that the Volterra property is instantiated in the target system. Our confidence that the Volterra property is instantiated in the target system might increase as a result of this to the extent that we believe that the assumption that the predator-prey system is negatively coupled is correct *and* also to the extent that we believe that there are no disrupting factors in the target system. But this should be seen as merely a possible *by-product* of the confirmatory power of robustness analysis.

[8]Kuorikoski et al. (2010) use the notation V_i to refer to tractability idealizations. To be consistent with my notation I have replaced all instances of V_i with T_i.

[9]In the above quote, Kuorikoski et al. do not explicitly claim that the models' results must also be probabilistically independent conditional on $\neg R_T$. But without this assumption this argument is not valid, so I am assuming this is just a slip of the hand.

To adequately assess this argument, I will need to make a couple of clarifications. First, as mentioned in Section 4.1, according to Kuorikoski et al. tractability idealizations are not the only kind of idealizations typically needed for the process of inference to be feasible; various Galilean idealizations will also be needed. So to be a little more rigorous one should say that a modeling result R can be derived from $C\&T_i\&G_i$ rather than just $C\&T_i$. But given that according to Kuorikoski et al., Galilean assumptions 'serve to isolate the working of the core causal mechanism by idealising away the influence of the confounding factors', I will assume for the sake of argument that Galilean assumptions, rather than being problematic, are always helpful in establishing causal dependencies. Therefore, I will assume that each model involves the same Galilean assumptions and I will set them aside for the time being.

Second, Kuorikoski et al. (2010, 545) reference Bovens and Hartman (2003, 96-97) to justify that the sort of probabilistic independence invoked in this argument is enough to guarantee that our degrees of belief in the hypothesis h should rationally increase. Indeed, Bovens and Hartmann (2003, 96-97) do show that under certain *specific* conditions, if distinct instruments' results are probabilistically independent conditional on the assumption that the testable consequence of a hypothesis is correct (or not correct),[10] then observing multiple positive results from distinct instruments should increase our degrees of belief in that hypothesis. But Bovens and Hartmann's demonstration depends on several *other* conditions being satisfied! One of these, for instance, is that an unreliable instrument must 'randomize at some level a':

> Our model does not apply to unreliable instruments that do not randomize, but rather provide accurate measurements of other features than the features they are supposed to measure. In effect, our model exploits the coherence of the reports as an indicator that the reports are obtained from reliable rather than unreliable instruments. But if unreliable instruments accurately measure features other than the

[10] Bovens and Hartman's definition of a testable consequence of a hypothesis is as follows: 'the probability of the consequence given that the hypothesis is true is greater than the probability of the consequence given that the hypothesis is false' (ibid. 90).

> ones they are supposed to measure, then they will also provide coherent reports and so the coherence of the report is no longer an indicator that they were obtained from reliable instruments. (Bovens and Hartmann 2003, 95)

So if Kuorikoski et al. want to appeal to Bovens and Hartman's demonstration to justify the validly their argument, then they also *must* rely on the assumption that unreliable models (in contrast to reliable ones) do not tend to give coherent reports. In other words, they must rely on the assumption that unreliable models cannot be systematically biased. As mentioned earlier, for simple measurement devices like scales this seems to be an adequate assumption in some cases: for any unreliable scale (i.e. malfunctioning) from a different manufacture and supplier it can be reasonable to assume that whether or not it shows that I weigh 300 pounds (if I really weight 300 pounds) is a matter of chance (and the same if I do not really weigh 300 pounds). But in the case of models (sharing the same substantial assumptions C, but differing in their tractability assumptions T_i), this is a *substantial* assumption that would need to be further justified and nowhere in the paper do Kuorikoski et al. do so. The fact that the validity of this argument depends on substantial assumptions, that have not been made explicit by Kuorikoski et al., is in my view already an important weakness of the argument, one that is possibly strong enough to reject it. But for the sake of argument, in this section I am going to assume this argument is valid and hence I will only critically assess whether, if valid, it is also sound.

For this argument to be sound Kuorikoski et al. need to convince us that it is reasonable to suppose that the probabilities of the models' results are independent conditional on R_T (and $\neg R_T$). But although this is a weaker notion of probabilistic independence than unconditional probabilistic independence, it is still an extremely strong notion of independence. This sort of independence demands that if I know that the models' result R is instantiated in the target system, then learning that a model gives result R should not at all affect my degrees of belief that another model would also give result R. But this is an unreasonable demand! To see why this is, suppose that I know that R is instantiated in the

target system. If this is all I know, then there is no reason to think that prior to learning the models' results, I will have much confidence in the fact that result R will be derived by these models (even if I know that C is instantiated in the target system as I have no reason to suppose that R causally depends on C!). But now suppose that I learn that R can be derived by one of the models, consisting of substantial assumptions and tractability assumptions $C\&T_1$. Kuorikoski et al.'s notion of independence demands that my degrees of belief about whether R can be derived from another model $C\&T_2$ should not change. But this is implausible: I know that the two models share substantial assumptions C, so if I learn that R can be derived from the first model, my degrees of belief about whether R will be derived by the second model are bound to change: I now seem to be in a much better position than I was before to make an informed guess that result R will be derived by second model.

To make my objection more vivid, consider the Lotka-Volterra model alongside another model which shares the substantial assumption C that the system is negatively coupled, but that involves a different set of tractability assumptions. Suppose that all I know is that the Volterra Property is instantiated in the target system. Given that I have no knowledge regarding what the Volterra Property causally depends on, there is no reason to suppose that I should have much confidence in the fact that the Volterra property will be derived by these two models. But now suppose that I learn that the Lotka-Volterra model has the Volterra property. Surely my degrees of belief that another model sharing the same substantial assumption C will also give the Volterra property will greatly increase. Why? Because the two models share substantial assumption C, and hence the fact that the first model had the Volterra property when assumption C was involved will greatly increase my confidence that the second model will also give the Volterra property.

Notice that the situation in my scale example is very different. Conditional on the fact that I really weigh 300 pounds, it seems reasonable to suppose that learning that a scale shows that I weigh 300 pounds will not affect my degrees of belief that another distinct scale will also show that I weigh 300 pounds. Effectively the difference consists in the following. Learning that a scale shows my

weight does not affect my degrees of belief that another scale will also show my weight, because I already knew that scales are supposed to measure my weight *prior* to learning the first scale's reading. Whereas in the case of models, the situation is very different: If all I know is that R_T is true, learning that a model with substantial assumption C gives result R will affect my degrees of belief that another model sharing substantial assumptions C will also give result R, because learning that the first model gives result R when C is involved, gives me some reasons to expect that the second model, which also involves C, will also give result R; reasons that I didn't have prior to learning the first model's result.

It is worth mentioning that Schupbach (2018) has also objected to this notion of conditional independence in the context of model-based robustness analysis, but his objection relies on the assumption that the distinct models will share many unrealistic assumptions and so 'discovering that one of the models is unreliable should often greatly increase our confidence that the other is too' (ibid., 283). In other words his objection is the following: conditional on the result R not being correct, the probabilities of the models' results R cannot be independent. This is indeed a very good objection, but it is weaker than mine because it relies on the idea that models will invariably share many unrealistic assumptions. Although this is certainly true in most if not all cases, the reason why I object to this notion of conditional independence is because of the very fact that the distinct models share substantial assumptions C and so it will hold regardless of whether or not they share any unrealistic assumptions.

All said and done, it seems to me that models cannot be independent in the way required by Kuorikoski et al.'s argument. Hence this argument is not sound and should be rejected.

4.2.3 A prima-facie more plausible argument (and yet...)

It is worth noting that alternatively to Bovens and Hartman's demonstration, Kuorikoski et al. might want to appeal to Fitelson (2001)'s demonstration instead. Fitelson (2001) shows that with respects to several popular Bayesian measures of confirmation, if two results a_1 and a_2 individually confirm an hypothesis H and if a_1 and a_2 are confirmationally independent regarding H, i.e.

$c(H, a_1|a_2) = c(H, a_1)$ and $c(H, a_2|a_1) = c(H, a_2)$, then a_1 and a_2 together confirm H to a greater extent than either a_1 or a_2 does separately, i.e. $c(H, a_2\&a_1) > c(H, a_1)$ and $c(H, a_2\&a_1) > c(H, a_2)$.[11] Fitelson further *suggests* that a sufficient condition for a_1 and a_2 to be confirmationally independent regarding H is that they be probabilistically independent conditional on H (and $\neg H$), i.e. $Pr(a_1\&a_2|H) = Pr(a_1|H)Pr(a_2|H)$ and $Pr(a_1\&a_2|\neg H) = Pr(a_1|\neg H)Pr(a_2|\neg H)$.[12] If this is right then Kuorikoski et al. could rely on Fitelson's 'result' but only if they are willing to change the notion of conditional independence they demand on the models' results. That is if they want to rely on Fitelson's demonstration then their argument should be rephrased as follows:

> **A second argument**. Assume that we observe that a model with substantial assumption C and tractability assumptions T_1 gives result R. Then we will have some degrees of belief that the hypothesis h: "In the actual world, R causally depends on C" is true. Suppose further that in addition to our first model, we observe that several other models sharing the same substantial assumptions C, but differing in their tractability assumptions T_i give the same result R. This should rationally increase our degrees of belief in the hypothesis h, because it is reasonable to assume that the models' results are probabilistically independent conditional on the hypothesis h (and $\neg h$).

This argument strictly relies on the assumption that each model's result individually confirms h. This does not seem to be an unreasonable assumption in most cases, but it is still an assumption that needs to be acknowledged.[13]

For this argument to be justified Kuorikoski et al. need to convince us that it is reasonable to suppose that the probabilities of the models' results are independent conditional on the hypothesis h (and $\neg h$). For instance, in the case

[11] A confirmation measure $c(H, a)$ measures the degree of confirmation lent to H by a. I use the notation $c(H, a_i|a_j)$ to indicate the degree of confirmation lent to H by a_i, conditional on a_j.

[12] To the best of my knowledge, however, Fitelson (2001) does not actually prove this result.

[13] Although someone may very well question this assumption too: if all the models involve false assumptions, why would we have to accept that there is any confirmation relation at all? I.e. why should we think that $c(H, R_1), c(H, R_2)$ etc. ... are not all equal to zero? Indeed, if they are all equal to zero, the machinery does not get off the ground.

of the Volterra principle, we want our models' results to be probabilistically independent conditional on the *Volterra principle*, rather than conditional on the *Volterra property* as in Kuorikoski et al's original argument. This kind of conditional probabilistic independence of the models' results is prima facie more plausible: conditional on the hypothesis that in the actual world R depends on C, the fact that two models (consisting of $C\&T_1$ and $C\&T_2$ respectively) share substantial assumptions C seems less of a salient factor when assessing one's degrees of belief that one model will give R if one has learnt that another model has already given R. To see why this is, suppose that I know that in the actual world R causally depends on C (e.g. I know that the Volterra principle is correct). In this case it seems that already *prior* to learning the models' results, my confidence in the fact that R will be derived by these models is going to be relatively high, since I know that they both involve C. That is, in this case knowing that R causally depends on C seems to already put me in a good position to make an informed guess that result R will be derived by both models. But then in this case, learning that one model gives R does not seem to put me in a better position to make an informed guess about whether the second model will also give R. Hence it seems, prima facie, plausible to assume that if I know that in the actual world R causally depends on C, learning that the model consisting of $C\&T_1$ gives R should not change my degrees of belief that a model consisting of $C\&T_2$ will give R.

However, despite this prima-facie plausibility, this sort of independence is still unrealistically strong in most cases, if not all. And I see two reasons for this. First, despite differing in *some* tractability assumptions, models will more often than not share many other tractability assumptions. But then, in these cases, if I learn that result R can be derived from the first model, it is unreasonable to suppose that my degrees of belief that R will be derived by the second model are not going to change: even if I know that R causally depends on C in the actual world, if the second model shares some tractability assumptions with the first model, learning that the first model gives R *will* put me in a better position to make an informed guess about whether the second model will also give result

R.[14]

So it seems that the only scenario in which it might be reasonable to assume that models' results satisfy this sort of independence is in those rare cases in which models share the same substantial assumptions C, but share no tractability assumptions. As far as the Lotka-Volterra model is concerned, Kuorikoski et al. (2012) argue that Weisberg and Reisman's (2008) individual based model - in which the Lotka Volterra model's variables, parameters and other assumptions are all translated into individual-based terms (an instance of what Weisberg and Reisman (2008) call representational robustness, which was discussed in section 3.3.1)- is one such case:

> Weisberg and Reisman (2008) also discuss a way in which practically all the tractability assumptions can be expected to be independent: the derivation of the robust theorem in a completely different modelling framework. Whereas the class of Lotka–Volterra models described above are sets of differential equations relating population aggregates, the Volterra principle can also be demonstrated using agent based computational models. Such models represent the same core causal mechanisms, albeit describing them at an individual level. However, the radical difference in the modelling framework means that the tractability assumptions, although still unavoidable, are of an altogether different kind: they relate to the behavioural rules of individuals and the spatial representation of their environment, rather than to population-level generalisations as in the original Lotka–Volterra models. (Kuorikoski et al., 2012)

If cases of representational robustness really are cases in which models share the same substantial assumptions C, but differ in *all* their tractability assumptions, as Kuorikoski et al. claim, then perhaps the sort of independence invoked in this argument is plausible in such cases. But notice that, if cases of representational robustness are the only kind of cases in which the sort of independence invoked

[14]Schupbach (2018, 285) makes essentially the same objection.

in this argument is plausible (as I am suggesting) then the scope of this argument is clearly very restricted. Hence this argument is not applicable in most instances of robustness analysis that are encountered in scientific and economic modelling.

But I think there is a second reason to doubt that this sort of independence is reasonable, even in the rare cases where models share the same substantial assumptions C but differ in all their tractability assumptions. And it has to do with the very nature of tractability assumptions. As mentioned previously, for Kuorikoski et al., tractability assumptions are assumptions that are introduced 'only for reasons of mathematical tractability' and, in contrast to Galilean assumptions, they often 'have no empirical merit on their own'. According to Kuorikoski et al. (2012), as far as the Lotka-Volterra model is concerned, an example of a tractability assumption is the *specific* functional form used to describe the rate of prey capture per predator, since *any* assumed functional form for the rate of prey capture will 'strictly speaking be false for any natural population' (ibid., 8). But although it is true that any assumed functional form will strictly speaking be false for any real-world predator-prey system, there is certainly a sense in which one particular functional form might be more adequate to describe the rate of prey capture than another, despite both of them being strictly false. And there is also a sense in which one might believe that at most one functional form amongst the ones one is considering is adequate, even if one lacks the knowledge to determine which one. But then, to the extent that this is the case, I think it is unreasonable to assume that the results of two distinct models that differ in their tractability assumptions are probabilistically independent conditional on h. And here is why. Suppose that I know that in the actual world R causally depends on C (e.g. I know that the Volterra principle is correct) and consider two distinct models that share substantial assumption C (e.g the assumption that the predator-prey system is negatively coupled) but that assume distinct functional forms for the rate of prey capture per predator. Suppose further that I believe that at most one of these two functional forms can adequately represent the actual rate of prey capture per predator. Prior to learning the models' results, I will have some degrees of belief in the fact that R (e.g. the Volterra

property) will be derived by these models. But now suppose that I learn that one of these models gives result *R*. This should give me further reasons to suppose that the particular functional form assumed in this model can adequately describe the rate of prey capture per predator, reasons that I didn't have prior to learning the model's result. And if that's the case, then this should also give me further reasons to suppose that the functional form assumed in the other model is inadequate. But then, to the extent that this is the case, it is unreasonable to suppose that learning that the first model gives result *R* will not change my degrees of belief that the second model will give result *R*. I now seem to have some further reasons to suppose that the second model does not adequately represent predator-prey systems, which should reasonably decrease my degrees of belief that the second model will give result *R*. Hence, it is hard to see why it would be reasonable to assume that conditional on hypothesis *h* being correct, these two models' results are probabilistically independent.[15]

Hence, due to the fact that in most cases of robustness analysis models will very often share many tractability assumptions, and due to the very nature of at least some tractability assumptions, I think it is in fact rather hard to justify the sort of probabilistic independence invoked in this argument in most if not all cases of model-based robustness analysis.

Before concluding, I would like to make one last remark. Throughout this section, I have assumed that there is a clear distinction between Galilean assumptions on the one hand and tractability assumptions on the other. In particular, I assumed that Galilean assumptions are always helpful in establishing causal dependencies by idealizing away the influence of the confounding factors. This allowed me to assume that robustness failure in a modelling result with respect

[15]Notice further, that for this argument to be applicable, there must be a clear conceptual difference between tractability assumptions on the one hand and Galilean assumptions on the other. In particular, models that involve different tractability assumptions must never describe systems which include distinct causal factors that can preempt the (alleged) stable capacity of the common core *C* to produce *R* from being manifested. This is because, as mentioned earlier, whether or not a causal factor may preempt a stable capacity from being manifested is an additional hypothesis that is independent of the truth of a robust theorem (e.g. the Volterra principle). Only under this assumption, it reasonable to assume (as I have in this section) that models that involve different tractability assumptions, but share the same substantial and Galilean assumptions can be used to confirm the *same* hypothesis. But if this assumption is unwarranted for the kind of assumptions that Kuorikoski et al.'s refer to as tractability assumptions, then all the worse for this argument, because inapplicable.

to Galilean assumptions is never epistemically problematic, since it merely suggests a new empirical hypothesis about a causally relevant feature. Without this assumption it would have been impossible to even begin to assess Kuorikoski et al.'s argument for the epistemic import of robustness analysis. This is because this assumption allowed me to give an *empirical* (causal) and unambiguous interpretation to the robust theorem. In other words, with this assumption I was able to interpret the robust theorem as a *causal* hypothesis about the *real* world, a hypothesis that one can both conditionalise on and confirm. However this distinction is, in my view, a lot less clear than Kuorikoski et al. suggest. Take the assumption that predators can consume infinite quantities of prey. This is arguably a Galilean assumption, since it assumes that there is no factor (e.g. a biological factor) that affects predator satiation. But a Volterra principle that only applies to target systems in which predators can consume infinite amount of food is clearly not a principle about real-world predators since no *real* predator can consume infinite amounts of food! This may not seem problematic under the assumption that if negatively coupled (fictional) predator-prey systems with no saturation have the *capacity* to produce the Volterra property then so must negatively-coupled predator-prey systems with saturation, despite the fact that their capacity may not be manifested. If that were the case, then learning that negatively-coupled predator-prey systems with no saturation have the capacity to produce the Volterra property would effectively be learning that predator-prey systems with saturation also have that capacity (and hence we would be learning something about real-world predator-prey systems). However, this assumption is in, my view, wrong. For if indeed we were to find out that negatively-coupled predator-prey systems with saturation didn't have the Volterra property,[16] what this would mean is not that their capacity to produce the Volterra property is not manifested, but rather what it would mean is that negatively-coupled predator-prey systems with saturation *no longer* have the capacity to produce the Volterra property. In other words, what we would learn in this case is that, like birds lose

[16]This is just an example to illustrate my point. Indeed, according to Weisberg (2006, 736) the Volterra principle still holds when we add a term for predator satiation.

their ability to fly because their wings are broken, negatively coupled predator-prey systems lose their capacity to produce the Volterra property because predators can't consume infinite quantities of prey. But if this is right, then it seems to me that, at least as far as *some* Galilean assumptions are concerned, if they are not de-idealised from the model, then no matter how many different sets of tractability assumptions we might go through, the theorem that we are actually trying to confirm does not seem to be a theorem about the actual world, but a fictional one. This does not necessarily mean that we can't conditional on this theorem (as required by this argument), and thereby confirm it, but it does raise the question as to what is the relevance of learning this theorem for learning about the real-world.

In this section, I reconstructed and critically assessed Kuorikoski et al.'s argument for the epistemic import of model-based robustness analysis. In Section 4.2.1, I argued that a causal argument from coincidence for the epistemic import of model-based robustness analysis is misleading and should be rejected. In Section 4.2.2, I tried to reconstruct what I take to be Kuorikoski et al.'s 'official' argument for the epistemic import of robustness analysis; I first argued that the validity of this argument relies on substantial assumptions that have not been made explicit by Kuorikoski et al. I then argued that, even if its validity is not brought into question, Kuorikoski et al.'s argument is not sound, since the sort of probabilistic independence on which it relies is unfeasible in *all* cases of robustness analysis. In Section 4.2.3, by revising the notion of probabilistic independence imposed on the models' results, I introduced a prima-facie more plausible argument for the epistemic import of robustness analysis. However, despite this prima-facie plausibility, I argued that it is in fact very hard to justify its soundness in most, if not all, cases of model-based robustness.

As mentioned at the beginning of this section, Odenbaugh and Alexandranova (2011) have also objected to Kuorikoski et al.'s argument for the epistemic import of robustness analysis. According to them, 'robustness analysis crucially depends on showing that the assumptions of different models are independent of one another'; one of their objection to Kuorikoski et al.'s argument is that 'reports of their independence have been greatly exaggerated' (ibid., 759). But

this objection suggests that the independence on which Kuorikoski et al.'s argument relies merely fails *in practice*, rather than *in principle*; this made it easy for Kuorikoski et al. (2012) to dismiss this objection - their argument relatively unharmed. Whereas I hope to have convinced the reader more forcefully that arguments that rely on some sort of probabilistic independence to justify the epistemic import of robustness analysis are implausible in most, if not all instances of robustness analysis. In particular, I hope to have shown that it is a mistake to assume that models might behave a bit like measuring instruments merely because this seems to fit well with our unquestioned intuitions. In other words, I hope to have shown that in our attempt to understand if, and when, looking at more than one model of the same phenomenon can help us learn about the world, we must, here as ever, rigorously question our intuitions rather than letting them dictate the kind of assumptions we are willing to accept.

In the next section, I will introduce Schupbach's (2018) recent Bayesian account of robustness analysis as explanatory reasoning and I will investigate under what conditions this account can be successfully applied to model-based RAs.

4.3 A critical assessment of Schupbach's explanatory account of model-based robustness analysis

As argued in the previous section, arguments that rely on some sort of probabilistic independence to justify the epistemic import of (model-based) robustness analysis are implausible in most, if not all instances of robustness analysis. However, this does not mean that the Bayesian can't rely on *other* arguments to justify its epistemic import, arguments that do *not* rely on probabilistic independence. The aim of this section is to critically assess one such argument recently offered by Schupbach (2018).

Like me, Schupbach (2018) also thinks that Bayesian accounts of robustness analysis (RA)[17] which rely on probabilistic independence to explicate the notion

[17] At this point a bit of terminological housekeeping is in order. The term 'robustness analysis' is an unfortunate one as it can mean different things to different people. As we have seen in Section 3.3, Weisberg (2006), uses the term to refer to a four-step procedure 'which begins by examining

of evidence diversity are in many cases, no matter how subtly formulated, woefully inadequate.[18] But if we are right it seems that in order to capture those cases the Bayesian *must* depart from independence-based accounts of RA diversity. Schupbach's (2018) recent explanatory account of RA has been rightly welcomed as a promising step in the right direction. Indeed, by having 'as its central notions explanation and elimination,' (ibid., 286) this account seems to fit very nicely with many empirically driven cases of RA in science, thereby revealing why these cases are able to lend confirmation to a hypothesis.

Schupbach, however, has further suggested that this explanatory account of RA 'applies to *model-based* RA just as well as it does to empirically driven RAs' (ibid., 297, my emphasis). The core aim of this section is to demonstrate that applying this account in the context of models is a lot more difficult than Schupbach suggests. The structure of this section is as follows. In Section 4.3.1, I will introduce Schupbach's explanatory account of RA. In Section 4.3.2, I will give an example of an empirically driven case of robustness analysis to illustrate how and why Schupbach's account can be successfully applied to this case. In Section 4.3.3, I will attempt to apply Schupbach's account to a case where the hypothesis we want to confirm through model-based RA is the Volterra principle (an instance of a 'robust theorem'). I will argue that although the application of Schupbach's account to model-based RA relies on several non-trivial assumptions, they may be reasonable in this case.

a group of similar, but distinct, models for a robust behavior and ends with the formulation of a robust theorem' (ibid., 737) and many have followed suit. However, in line with Schupbach's notation, in this section I am using the term 'robustness analysis' more broadly than Weisberg does, to refer to the general practice of using multiple means to detect the same result, where those means 'could include experiments, laboratory instruments, sensory modalities, derivations (from axioms, models, theories, and so on), axiomatic systems, computer simulations, and formal models amongst other things' (Schupbach 2018, 277). However, this is purely a terminological decision, and with this decision I am not at all suggesting that there aren't in fact distinct kinds of RA with important differences between them, differences which have implications for our assumptions and their epistemic import (as has been argued by several philosophers; see for instance Woodward (2006), Calcott (2011)).

[18] Schupbach (2018) considers three accounts of probabilistic independence that could be used to explicate the notion of evidence diversity: unconditional probabilistic independence, reliability independence and confirmational independence (the latter two were discussed in the previous section). And he argues that in what he considers some paradigmatic cases of RA in science the assumptions on which these accounts rely are implausible. I think there is scope for disagreement as to whether the cases he considers really are *paradigmatic* cases of RA in science. However, as long as RA is understood as the general practice of using different means to detect the same result, the cases he considers count as clear cases of RA in science whether they are paradigmatic or not.

4.3.1 Schupbach's explanatory account of RA diversity

According to Schupbach, when there is more than one means of detecting a result R, the notion of diversity that is relevant to RA is the following:

> **ERA Diversity**:[19] Means of detecting R are ERA diverse with respect to potential explanation (target hypothesis) H and its competitors to the extent that their detections $(R_1, R_2, \ldots, R_n)$ can be put into a sequence for which any member is explanatorily discriminating between H and some competing explanation(s) not yet ruled out by the prior members of that sequence. (Schupbach 2018, 288)

Of course, the above account of ERA diversity would leave too many questions unanswered: what counts as a potential competing explanation of R? What does it take for a detection of a result R by a given means to be explanatorily discriminating between the target hypothesis H and some competing explanation H'? And why should one consider ERA diversity to be epistemically important from a Bayesian perspective? By relying on a probabilistic conception of explanatory power $\varepsilon(H, E)$, Schupbach attempts to provide an answer to these questions.

According to Schupbach, the explanatory power an explanation H has over its explanandum E is given by

$$\varepsilon(E, H) = \frac{Pr(H|E) - Pr(H|\neg E)}{Pr(H|E) + Pr(H|\neg E)}, \tag{4.1}$$

where $\varepsilon(E, H)$ can take values ranging from $[-1, 1]$ and the greater the value of $\varepsilon(E, H)$, the more strongly H explains E.[20] Since $\varepsilon(E, H) = -\varepsilon(\neg E, H)$, this

[19] Contrary to what Schupbach (2018) seems to suggest, I don't believe this to be the only notion of diversity that is relevant to RA from a Bayesian perspective. For although Schupbach convincingly argues that accounts which rely on probabilistic independence to explicate the notion of evidence diversity are implausible in the two cases of RA which he uses to motivate his explanatory account of RA, this is far from showing that those accounts are implausible in *all* cases of RA in science (and it is in my view misleading to suggest otherwise). Hence what he calls **RA diversity**, I will call Explanatory Robustness Analysis diversity or **ERA diversity**.

[20] Schupbach and Sprenger (2011, 107) are careful in pointing out that this probabilistic measure of explanatory power 'is not intended to reveal the conditions under which a theory is explanatory of some proposition [...]; rather, its goal is to reveal, for any theory already known to provide such an explanation, just how strong that explanation is.' Although a variety of distinct probabilistic measures of explanatory power have been proposed in the literature, for the purpose of what I will be arguing in this section, I am happy to assume that this measure does a good job of capturing the explanatory power that an explanation has over its explanandum (for a defence of this probabilistic measure of explanatory power see Schupbach and Sprenger (2011) and for

means that if $\varepsilon(E,H) > 0$, H explains E more strongly than it explains its negation; if $\varepsilon(E,H) < 0$, H explains E less strongly than it explains its negation; and if $\varepsilon(E,H) = 0$, H is explanatory irrelevant to E.

It can be shown that the value of $\varepsilon(E,H)$ is positively correlated with the degree of statistical relevance between E and H, that is, the strength of the inequality $Pr(E) < Pr(E|H)$ (see Schupbach and Sprenger (2011, 110)). Hence, according to this measure of explanatory power, the more H decreases the degree to which E is surprising, the more strongly H explains E. Below are some additional properties of $\varepsilon(E,H)$ that show how the impact that H has on the degree to which E is surprising is related to the explanatory power that H has over E. As long as $Pr(H) \not\approx 0$ and $P(E) \not\approx 0, 1$, then:

- $\varepsilon(E,H) > 0$ iff $Pr(E|H) > Pr(E)$ and $\varepsilon(E,H) < 0$ iff $Pr(E|H) < Pr(E)$;
- $\varepsilon(E,H) \approx 1$ iff $Pr(E|H) \approx 1$ and $\varepsilon(E,H) \approx -1$ iff $Pr(E|H) \approx 0$;
- $\varepsilon(E,H) = 0$ iff $Pr(E|H) = Pr(E)$.

Similarly, the explanatory power that an explanation H has over its explanandum E, in light of some proposition p, is given by

$$\varepsilon(E,H|p) = \frac{Pr(H|E\&p) - Pr(H|\neg E\&p)}{Pr(H|E\&p) + Pr(H|\neg E\&p)}. \tag{4.2}$$

Equipped with this probabilistic conception of explanatory power, Schupbach (2018, 293) provides the following five formal conditions for a successful increment of ERA diversity:

> **Past detections**: We are given $E = R_1\&R_2\&\ldots\&R_{n-1}$ (informally, a result R has been detected using $n-1$ different means);
>
> **Success**: $\varepsilon(E,H), \varepsilon(E,H') > 0$ (informally, the target hypothesis H explains this coincidence, but so does another rival hypothesis H');

some criticisms see Glymour (2014)). However, it is worth mentioning that Schupbach (2018, 292) claims that all of the substantive results derived using this measure in his article also hold using any of the alternative measures defended in Popper (1959), Good (1960), McGrew (2003) and Crupi and Tentori (2012).

Competition:[21] (i) $Pr(H\&H') = 0$, or (ii) $\varepsilon(E, H|H') \leq 0$) (informally, H and H' epistemically compete with one another, with respect to E);

Discrimination: $\varepsilon(R_n, H|E) \approx 1$, $\varepsilon(\neg R_n, H'|E) \approx 1$ (informally, there is another nth means of potentially detecting R such that, in light of E, H would strongly explain the detecting of R by this means (R_n) and H' would strongly explain not detecting R by this means ($\neg R_n$));

New detection: we learn R_n (informally, the nth means also detects result R).

Schupbach then shows that the above formal conditions, if satisfied, *guarantee* an incremental confirmation of the target hypothesis H, i.e. $Pr(H|E\&R_n) > Pr(H|E)$ (for a proof see Schupbach (2018, 293-296)). In other words, evidence that is ERA diverse[22] with respect to a target hypothesis H and a competing hypothesis H' must incrementally confirm H.[23]

In light of these formal conditions and of what those conditions entail, it is now clear why Schupbach's notion of ERA diversity is epistemically important from a Bayesian perspective: evidence that is ERA diverse with respect to a target hypothesis H and its competitors should rationally increase one's degrees of belief in that hypothesis. Clearly, however, this fact alone says nothing about how *much* one's confidence in H should increase: the increase warranted by the evidence could be anything from negligible to substantial. Luckily Schupbach also has something to say about the extent to which a successful increment of ERA diversity should increase one's degrees of belief in H. The answer, however, will depend on whether or not the target hypothesis H and the competing hypothesis H' are mutually exclusive (i.e. on whether case (i) or case (ii) in the

[21] According to Schupbach, there are two ways for hypotheses to epistemically compete. Two hypotheses H and H' might epistemically compete because they are mutually exclusive, i.e. case (i). But they might also epistemically compete because H' suffices to do the explanatory work of H, i.e. case (ii).

[22] The above conditions for a successful increment of ERA diversity also define ERA diversity itself, since these are the conditions that have to be satisfied for two distinct means of detecting a result R to count as ERA diverse.

[23] Notice that neither the success condition nor the discrimination conditions can be satisfied if either $Pr(H) = 0$ or $Pr(H') = 0$. So if an agent has no confidence in either H or H' there cannot be any successful increment of ERA diversity.

competition condition is satisfied). When H and H' are mutually exclusive, Schupbach shows that the lower bound of how much H is confirmed by a successful increment in ERA diversity is determined by the agent's prior degrees of belief in H', $Pr(H')$, and by the likelihood of H' on E, $Pr(E|H')$. The higher $Pr(H')$ and $Pr(E|H')$ are, the higher that lower bound.[24] This result is very helpful. For although it doesn't give us the exact increment of confidence in H that is warranted by a successful increment of ERA diversity, it does nonetheless give use a lower bound on what that increment should be. When H and H' are not mutually exclusive what determines the extent of confirmation of H is more complicated, and given that the details don't matter for this section, I refer the reader to Schupbach (2018, 295-296) for a discussion of this case.

But what should one make of the above ERA diversity conditions? Are they intuitive? Do they fit nicely with actual cases of RA in science?

4.3.2 Empirically driven RAs

To motivate the intuition behind these conditions, Schupbach considers the case of the at the time curious motion of a sample of pollen granules suspended in water, first observed in 1827 by the botanist Robert Brown. In the early 20th century, Einstein famously offered an explanation for this observation: this motion, according to Einstein, was due to random molecular collisions in the water. Later, Jean Perrin performed a variety of experiments to determine if Einstein's molecular explanation for this motion (nowadays known as Brownian motion) was correct.[25] As Schupbach points out, the fact that this motion had been detected by a multitude of other different experiments (using different materials, different media, different means of suspending the particle, etc.) was considered

[24]The lower bound is determined by $Pr(H')$ and $Pr(E|H')$ since

$$\frac{Pr(H|E\&R_n)}{Pr(H|E)} \geq 1 + \frac{Pr(H')Pr(E|H')}{c}, \tag{4.3}$$

for some constant $0 \geq c \leq 1 - Pr(H')$. So the higher $Pr(H')$ and $Pr(E|H')$ are, the higher the lower bound is. See Schupbach (2018, 294) for a proof.

[25]As Mayo (1996, 44) explains 'doing so was regarded as a test of the kinetic theory against the classical theory of thermodynamics. If Brownian motion could be explained as caused by something either outside the liquid medium or within the particles themselves, then it would not be in conflict with the classical theory. If, alternatively, the cause of Brownian motion was shown to be a molecular motion in the liquid medium, as given in the kinetic theory, it would be in conflict.'

by Perrin (1913, 83-86) as evidence in support of Einstein's explanation.[26] But why should the robustness of Brownian motion have counted as evidence for Einstein's molecular explanation?

According to Schupbach, this is because the various means of detecting the Brownian motion were ERA diverse with respect to Einstein's molecular explanation and its competitors. Indeed, when Brown first observed the curious motion of the pollen granules suspended in water (R_1), there were more than a few competing explanations for this observed phenomenon: the motion might have been due to currents or evaporation of the water, or it might have been due to a sexual drive inherent in pollen, etc. But there were many later detections of this motion that were able to explanatorily discriminate between Einstein's molecular explanation H and one of the many competing explanations not yet ruled out. Take, for instance, the competing explanation H' that the motion was due to a sexual drive inherent in pollen. And consider a new detection of this motion using an inorganic material (R_2). Does this new detection satisfy all the five conditions for a successful increment of ERA diversity? Let us go through each of them:

In this example, the Brownian motion has already been detected using a sample of pollen granules suspended in water so we have $E = R_1$ and hence the **past detection** condition is satisfied. Furthermore, Einstein's molecular explanation H and the sexual drive inherent in pollen explanation H' provide different causal explanations for the observed motion R_1, so it seems reasonable to assume that both H and H' increase the probability that that we should observe this motion (i.e. $Pr(R_1|H) > Pr(R_1)$ and $Pr(R_1|H') > Pr(R_1)$). But then, it is also reasonable to assume that $\varepsilon(R_1, H) > 0$ and $\varepsilon(R_1, H') > 0$ (see section 2). Hence both H and H' plausibly satisfy the **success condition**. The **competition condition** is also plausible. For although H and H' are not mutually

[26]This, however, was but a very small subset of experimental results that Perrin (1913) cites as evidence for Einstein's molecular explanation of Brownian motion. It is also worth mentioning that most philosophical discussion on the extent to which robustness reasoning played a role in Perrin's arguments for Einstein's molecular explanation focuses on the convergence of Perrin's estimations of Avogadro's number from a variety of experiments on numerous distinct phenomena (Brownian motion, blackbody radiation, the blueness of the sky, etc.). So it is somewhat surprising that Schupbach actively chooses not to discuss this case of RA in relation to his account of ERA diversity. For a philosophical discussion of what role this case of RA played in Perrin's arguments see, for instance, Mayo (1996), Psillos (2011), Chalmers (2011) and Hudson (2018).

exclusive hypotheses, H' seems sufficient for doing the explanatory work of H, i.e. $\varepsilon(E, H|H') \leq 0$). It also seems plausible to assume that in light of the detection of this motion using a sample of pollen granules R_1, H would strongly explain the detection of this motion using an inorganic material (R_2), whereas H' would strongly explain not detecting this motion by this means ($\neg R_2$), in accordance with the **discrimination condition**. Why? H cites causes of the observed motion that would also cause the movement of inorganic material, whereas H' cites causes that would *not* cause such movement. Hence, it seems plausible to assume that whereas H makes it extremely likely that we would observe this motion using an inorganic material (i.e. $Pr(R_2|H\&R_1) \approx 1$), H' makes it extremely likely that we would *not* observe it (i.e. $Pr(\neg R_2|H'\&R_1) \approx 1$).[27] And this implies that $\varepsilon(R_2, H|R_1) \approx 1$ and $\varepsilon(\neg R_2, H'|R_1) \approx 1$. Finally, the Brownian motion has been detected using inorganic material (i.e. we learn R_2) and hence the **new detection condition** is also satisfied.

All conditions of ERA diversity seem plausible in this example. A similar story could, arguably, be told for many other means that were used by Perrin to detect Brownian motion. So I am happy to be enticed by Schupbach into concluding that the reason why the robustness of Brownian motion across various different means both was, and should have, counted as evidence for Einstein's molecular explanation is that each new detection of this motion lead to a successful increment of ERA diversity and hence, for this reason, each detection was able to incrementally confirm it. I have not attempted to convince the reader that

[27] One might object: why doesn't the sexual drive inherent in pollen explanation merely fail to make it likely (rather than make it extremely unlikely) that we would observe this motion using an inorganic material? This is a fair objection. In response, one might argue that upon accepting H' one no longer has any reason to accept any other potential explanation of E and hence all the other potential explanations of E should be ruled out. And if this is so, then it does seem reasonable to assume that upon accepting H', it is extremely likely that we wouldn't observe this motion. But why should we dismiss all potential explanations of E upon accepting H'? The **competition condition** is certainly relevant here. If H and H' are mutually exclusive then we should of course rule out H upon accepting H'. And if H' suffices to do the explanatory work of H then E is already explained and hence 'the explanandum no longer compels us to hunt for, and reason to, further explanations' (Schupbach 2018, 291). But there are two problems with this response. First, as Schupbach himself acknowledges, even though H' suffices to do the explanatory work of H, 'there may remain explanatory reason apart from E still supporting [H]' (ibid. 291). Second, for this response to work, upon accepting H' one also has to dismiss *all* other potential explanations of E, not just H. And for this to be somewhat plausible it must be the case that H' also epistemically competes with each and every one of these explanations, not just H. But this is not entailed by the **competition condition**. Hence this assumption would have to be defended separately.

Schupbach's account of ERA diversity is an adequate account of RA diversity in general, not least because I don't think it is. But what matters for this section, is that this account seems to fit very nicely with *some* cases of empirically driven RA.[28] In particular, what allowed us to apply Schupbach's account to this case is that we were able to find both an adequate target and rival explanation for the two detections of Brownian motion that plausibly satisfied all of the conditions of ERA diversity.

Schupbach argues that this account of ERA diversity 'applies to model-based RAs just as well as it does to empirically driven RAs' (ibid., 297). However, in the next subsection, I will show that when it comes to model-based RA the picture is rather more complicated than he suggests. In particular, I will show that, in contrast to the case above, it is not at all straightforward to formulate an adequate target hypothesis and rival hypothesis that satisfy all the conditions for a successful increment of ERA diversity. Whether this is possible relies on several substantial assumptions, assumptions which may be reasonable in some cases, but certainly not in others.

4.3.3 Does Schupbach's account of ERA diversity apply to model-based RA?

Schupbach claims that his account of ERA diversity can finally give an adequate Bayesian justification for why model-based RA can increase our confidence in the Volterra principle:

> When seeking to confirm the Volterra principle, [ERA]-diverse models may be quite similar apart from some modest differences in their simplifying assumptions. But by utilizing these distinct (though perhaps overall quite similar) models, we may eliminate confounding explanations of our result left standing by either model used alone. [...], we may discard worries that our result is an artefact of a particular unrealistic assumption of the first model by using a second model that does not share that assumption. (Schupbach 2018, 289)

[28] Whereas, as Schupbach (2018) convincingly argues, accounts which rely on probabilistic independence to explicate the notion of evidence diversity do not fit nicely with this case.

But I will show that the picture is considerably more complicated than what he suggests.

There is a substantial difference between empirically driven RAs and model-based RAs and it is important to make this difference clear before we can attempt to apply Schupbach's account of ERA diversity to the latter. Recall that Schupbach's account of ERA diversity concerns distinct means of detecting the same result R. Schupbach has shown that if those distinct means of detecting a result R are ERA diverse with respect to a target explanation H and its rival explanations for their detections $R_1, R_2, \ldots, R_n$, then H is incrementally confirmed. In the empirically driven case of RA considered in section 3:

- R is Brownian motion in the *actual world*;
- R_i are the distinct detections of Brownian motion in the *actual world*;
- H is a hypothesis about why we detect Brownian motion in the *actual world* (i.e. Einstein's molecular explanation).

So, in this case, R and its detections $R_1, R_2, \ldots, R_n$ all concern the actual world and the hypothesis that we want to confirm (i.e. Einstein's molecular explanation), which also concerns the actual world, is a possible explanation of these detections.

In this case of RA, however, things are less straightforward, since we have:

- R is the Volterra property in *model land*;
- R_i are the detections of the Volterra property in *model land*;
- H is a hypothesis about why we detect the Volterra property in *model land*.

So, in this case, R and its detections $R_1, R_2, \ldots, R_n$ all concern model land and since the hypothesis that we want to confirm (i.e. the Volterra principle) concerns the actual world, it is *not* a possible explanation for these detections.

Hence a crucial difference between empirically driven RAs and model-based RAs is the following: in the former, the hypothesis that we want to confirm is a possible explanation for why we detect the same result, whereas in the latter

it is *not* a possible explanation for why we detect the same result. In light of this difference, it is clear that the application of Schupbach's account of ERA diversity to model-based RAs is considerably less straightforward. This does not imply that Schupbach's account is not applicable to model-based RAs, but it does nonetheless show that any attempt to successfully apply it will have to acknowledge this difference, and show that it can be applied in spite of it. In this section, I will attempt to do just this.

To assess whether Schupbach's account of ERA diversity can apply to this example of model-based RA, it will be helpful to consider a very simple case. Suppose I have already learnt that the original Lotka-Volterra model has the Volterra property (i.e. I have learnt R_1). Suppose further that I subsequently learn that another model in which an idealization/assumption A_1 of the original Lotka-Volterra model has been replaced by an other idealization/assumption A_2 also has the Volterra property (i.e. I learn R_2). Are these two detections of the Volterra property ERA diverse with respect to a target and rival hypothesis? Or in other words, can Schupbach's account of ERA diversity show that learning R_2 should incrementally confirm the Volterra principle?

For this to be the case, we must find an adequate target hypothesis H and rival hypothesis H'. Let's think first about a plausible candidate for H. Of course, the hypothesis that we ultimately want to confirm is the Volterra principle. But as mentioned earlier, the Volterra principle cannot be an adequate target hypothesis since it is not a possible explanation for why we detect the Volterra property in model land. Indeed, recall that the **Success condition** demands that $\varepsilon(R_1, H) > 0$ and the **Discrimination condition** that $\varepsilon(R_2, H|R_1) \approx 1$. But if we take H to be the Volterra principle, neither condition is satisfied, since this hypothesis alone, without any further assumption about the ability of the models to adequately represent the real behavior of predator-prey systems, doesn't make it any more or less likely that the the two models in question have the Volterra property and hence $\varepsilon(R_1, H) = 0$ and $\varepsilon(R_2, H|R_1) = 0$. So to satisfy both conditions, the target hypothesis H must assert something like the following:

> *H*: The Volterra principle is correct & both the Lotka-Volterra model and the new model adequately represent the target system.

But what does it mean for a model to adequately represent the target system? Before I explain this, it will be useful first to contrast my choice of target hypothesis with that of Schupbach. Here is what he writes:

> That a biological model behaves in accordance with the Volterra principle while not making the unrealistic assumption that prey cannot take cover is explained well by the Volterra principle itself (*in conjunction with the hypothesis that the model is accurately modelling the real world behaviour of predator–prey systems*); (Schupbach 2018, 288, my emphasis)

Clearly, Schupbach also thinks that the target hypothesis cannot merely be the Volterra principle. According to him, the target hypothesis must be a *conjunction* of two hypotheses: the hypothesis that the Volterra principle is correct; and the hypothesis that 'the model is *accurately* modelling the real world behaviour of predator-prey systems'. However, I think the word *accurately* is rather misleading here. As discussed Chapter 3, we already know that the Lotka-Volterra model is highly unrealistic in many respects; that is, we already *know* that the Lotka-Volterra model does not accurately describe the real-world behaviour of predator-prey systems. Hence Schupbach's hypothesis cannot be a plausible candidate for the target hypothesis, since we already know from the outset that it is not true.

But that the Lotka-Volterra model is not an accurate representation of the target system is no surprise since, as many have argued, this is the case for most if not all scientific models. But then, if the hypothesis that a model is an accurate representation is not the sort of hypothesis that can be confirmed, what hypothesis might we try to confirm instead? According to Parker (2009, 2020), it is the adequacy of a model for a particular purpose. Under this view, what we want to (and maybe can) confirm is the hypothesis that a model, despite not being an *accurate* representation, is nonetheless an *adequate* representation for the particular

purpose at hand.[29] Following Parker's suggestion, the claim that 'the Lotka-Volterra model and the new model adequately represent the target system' in my target hypothesis is meant to capture the idea that the models are adequate for the particular purpose at hand, in this case, that of discerning whether or not the Volterra principle is correct. So in this case, the models adequately represent the target system just in case they have the Volterra property iff the Volterra principle is correct. Notice that if H is true - and hence the Volterra principle is true and both models adequately represent the target system - then both models must have the Volterra property, that is $Pr(R_1|H) = 1$ and $Pr(R_2|R_1\&H) = 1$. This implies that $\varepsilon(R_1, H) = 1$ and $\varepsilon(R_2, H|R_1) \approx 1$ and hence the **success condition** and **discrimination condition** are satisfied.

Have we found a plausible candidate for the target hypothesis? Perhaps. But of course only if we think that it is plausible that both models can indeed be adequate representations of the target system. This is not a trivial assumption. For instance, this might fail to be a plausible assumption if by replacing the old assumption A_1 with A_2, we believe that the new model now describes a system in which a causal factor can preempt the capacity of a negatively-coupled predator-prey system to manifest the Volterra property. In this case it is unclear why we should think that the new model can be adequate for discerning whether the Volterra principle is true. In Cartwright's words, if a model is to teach us about capacities, it must do so by 'mimicking Galilean experiments' where a Galilean experiment is 'one that isolates the cause under study so that it operates "without impediment". What happens in the experiment then is the exercise of that capacity and of that capacity alone' (Cartwright 2009, 47). However, if we know that the new model does not describe such a system, then this might be a reasonable assumption.[30] It is also worth pointing out that whether or not H is a plausible target hypothesis in this example does not rely on the

[29] See Katzav (2014), Frisch (2015) and Parker (2020) for discussions of the challenges and considerations involved in the confirmation of adequacy-for-purpose hypotheses (in relation to climate models).

[30] This of course relies on the idea that there is always a clear distinction between factors that trigger or fail to trigger a capacity on the one hand, and factors that preempt or don't preempt a capacity from being manifested on the other. Although, in my view, this is not a trivial assumption, it is one that must be true if a conceptually coherent interpretation of 'robust theorems' in terms of stable capacities is possible.

idea that the new assumption is necessarily more realistic than the old one. For although this might, arguably, be what Schupbach has in mind in light of his choice of A_1 (i.e. the unrealistic assumption that prey cannot take cover) and A_2 (i.e. the more realistic assumption that prey can take cover), this doesn't have to be the case. For instance, A_1 could be the specific functional form used to describe the rate of prey capture per predator (the Lotka-Volterra model assumes that there is a linear increase in prey capture with prey density) and A_2 could be another functional form. And we might consider both of these functional forms to be equally unrealistic idealizations. However, we may nonetheless believe that both of these functional forms are sufficiently accurate descriptions of the rate of prey capture per predator for *some* real predator-prey system. Hence in this case we can safely assume that it is possible for both the Lotka-Volterra model and the new model to be adequate for discerning whether the Volterra principle is true.[31]

An important caveat is in order here. For H is to be a plausible candidate for the target hypothesis we *must* think that both the Lotka-Volterra model and the new model are adequate representations not by mere luck but because, for whatever reasons, the model 'latches on' to the underlying mechanism in the target system. For if we do think that the models are adequate simply as a matter of luck, then according to us H is an arbitrary conjunction: that is the hypothesis that the Volterra principle is true and the hypothesis that the models behave in accordance to it (and hence they are adequate) are unrelated to one another. But if H is an arbitrary conjunction, the hypothesis that the Volterra principle is true is *irrelevant* to the behaviour of the models and hence it is not part of an explanation for the models' results. However, one might wonder: in virtue of what does a model 'latch on' to the underlying mechanism? Although a rigorous answer to his question is beyond the scope of this section, I do believe that any such answer will have to necessarily depend on the nature of the hypothesis we want to confirm. For instance, since in this case the hypothesis we want to

[31] Crucially, the reason why the incompatibility of A_1 and A_2 is not problematic in this example is that we are *not* interested in learning something about a *specific* predator-prey system. However, as I will argue in the next chapter (section 5.3), in cases where we are interested in learning something about a specific target system, the incompatibility of assumptions will present a problem for the applicability of Schupbach's account.

confirm is (by assumption) a claim about stable capacities, we might think the models are adequate not by mere luck just in case they all *successfully* mimic a Galilean experiment. However, in cases where we want to confirm a different type of hypothesis, the answer to this question will inevitably be a different one. But in any case, it is important to recognize that for H to be a plausible candidate for the target hypothesis, we *must* think that there is some connection between the target system and the behaviour of the models, and that it is in virtue of this connection that the models are adequate representations.[32]

We may have found a plausible candidate for the target hypothesis H. What about a plausible candidate for the rival hypothesis H'? This is what Schupbach writes:

> [...] but the competing explanation that this behaviour is attributable to the unrealistic assumption in question would rather provide a strong explanation of our failing to observe the behaviour using such a model. Such a model thus explanatorily discriminates between these potential explanations. (Schupbach 2018, 288)

Schupbach is suggesting that a rival explanation for the Lotka-Volterra model's result R_1 (i.e. the Volterra property) and the new model's potential result $\neg R_2$ is the hypothesis that the model's behaviour 'is attributable to the unrealistic

[32]During conversations, two alternative candidates for a target hypothesis H have been suggested to me. Although I don't think these suggestions work, they are worth mentioning in case the reader wants to think about this further. One suggestion is that H can be the Volterra principle itself as long as we think of the Volterra principle as a hypothesis which concerns both the actual world and model land. However, assuming that a Volterra principle which concerns model land must be interpreted as something like 'all models in the relevant class of models have the Volterra property', this suggestion for H can't work since H is an arbitrary conjunction of two hypotheses: the hypothesis that the Volterra principle holds in the actual world & the hypothesis that the Volterra principle holds in model land. And the latter is supposedly doing all the 'explaining', not the former. The other suggestion is that H is the hypothesis that all models in the relevant class of models have the Volterra property in conjunction with the hypothesis that at least one of the models in this class is adequate. However, the hypothesis that the Volterra property is present in the relevant class of models is again an arbitrary conjunction of hypotheses (model 1 has the Volterra property & model 2 has the Volterra property & ...). And if this is so, then the fact that we detect the Volterra property in some models does not confirm the hypothesis that all models in the relevant class of models have the Volterra property. As Lange (2001, 577) remarks, 'a hypothesis believed to be coincidental if true, such as "All of the families on my block have two children." [...] does not have its predictive accuracy confirmed by its success in a given case. For example, that the Jones family on my block has two children typically fails to confirm that the Smith family on my block does, too.' The crucial difference between my suggestion for H and the ones above is that the hypothesis that the Volterra principle is true and the hypothesis that models we select for RA (that may indeed be instances of a larger class of models) are all adequate (not by mere luck) for the purpose at hand is not an arbitrary conjunction.

assumption in question.' But notice that since *all* assumptions and idealizations of a model are needed for the derivation of a model's result, what this hypothesis actually means, according to Schupbach, is not clear. However, in my view (and perhaps Schupbach's too), the only possible candidate for a rival explanation H' is the following *logical* hypothesis:

> H': the Lotka-Volterra model entails R_1 & if A_1 is replaced with a different assumption (i.e. A_2) the new model entails $\neg R_2$.

The target hypothesis H and the rival hypothesis H' are mutually exclusive, since H' entails $R_1 \& \neg R_2$ whereas H entails $R_1 \& R_2$; hence the **competition condition** is satisfied. The fact that H' entails both R_1 and $\neg R_2$ means that it also satisfies both the **success condition** and the **discrimination condition**. Hence, since H' satisfies all conditions of ERA diversity, whether H' is a plausible candidate for a rival potential explanation will ultimately depend on whether or not we think that logical hypotheses can be explanatory in the first place (as mentioned in footnote 3, ε is not supposed to reveal whether a theory is explanatory of some proposition).[33]

If logical hypotheses are explanatory then we may we have found both a target and rival explanation that satisfy all of Schupbach's conditions of ERA diversity and hence detections R_1 and R_2 could count as ERA diverse in this case. However, some qualifications are in order. Notice that a logically omniscient agent will either have degrees of belief $Pr(H') = 1$ or $Pr(H') = 0$ depending on whether H' is true or false respectively; and since H and H' are mutually exclusive this means that for such agent either $Pr(H) = 0$ or $Pr(H') = 0$

[33]One may wonder if a non-logical alternative rival explanation H' could be found instead. For instance, one may suggest the following alternative for H': 'the Volterra principle is not true & the original model is inadequate & the new model is adequate. However, I have two concerns about this alternative rival explanation. First, it is unclear why the hypothesis the Volterra principle is not true and that the original model is inadequate can be thought of an explanation for why the original model has the Volterra property (in other words, why should the fact that a model is inadequate explain a particular result of the model?). Worryingly, if that were right, what would stop us from formulating a *target* hypothesis which states that 'the Volterra principle is not true & both models are inadequate'? This would clearly be bad news. Second, whether or not the Volterra principle is true is irrelevant for coming up with a rival explanation to my target hypothesis. According to my target hypothesis both models must indicate the truth of the Volterra principle since they are both adequate and the Volterra principle is true. Hence if the second model fails to indicate its truth, then the target hypothesis is rejected, independently of whether Volterra principle is true or false. Hence given that the falsity of Volterra principle is irrelevant for rejecting the target hypothesis, it is not clear why it should be part of a rival explanation.

and hence detections R_1 and R_2 cannot count as ERA diverse (see footnote 23). Hence, only for an agent who is not logically omniscient, for whom $Pr(H')$ can take non-maximal values, could detections R_1 and R_2 count as ERA diverse.[34] Furthermore, notice that given that, as mentioned in Section 4.3.1, the extent to which the target hypothesis is confirmed is partly determined by how plausible the rival hypothesis is prior to elimination, the extent to which H will be confirmed in this case would have to partly depend on the agent's knowledge and beliefs about the derivational relationships in a family of models. Clearly, this can vary substantially from agent to agent. Hence, although non-omniscient agents might agree that detections $R_1\&R_2$ are ERA diverse, they might nonetheless strongly disagree about the extent to which this should confirm H. In other words, the extent to which H will be confirmed is highly contextual and, arguably, also very difficult to assess within a given context (since it requires an agent to assess their own knowledge and beliefs about the various derivational relationships in a family of models: evidently not an easy task). Finally, it is worth noting that there might be cases where there are an infinite number of idealizations with which we could replace a particular idealization (e.g. there is an infinite number of functional forms that we could pick to describe the rate of prey capture per predator). So one may wonder: when can we stop worrying about the infinite number of rival explanations that have not yet been ruled out by our finite number of ERA diverse detections? There are, however, two considerations that might help with this question. First, we might think that a large class of those idealizations are not sufficiently accurate for any predator-prey system hence whether or not a model which includes one of those idealizations has the Volterra property is irrelevant to us because we don't think that the model is adequate. Second, and perhaps one of the main lessons to take away from Schupbach's account of ERA diversity, only the elimination of the rival (logical) hypotheses that we think are plausible can incrementally confirm H. Hence if we strongly believe that by replacing an idealization with another,

[34] Although there have been various attempts to relax the logical omniscience assumption in Bayesian confirmation theory (see, for instance, Garber (1983)), not all Bayesians are willing to relax this assumption, as this move threatens to prevent the derivation of most important results in Bayesian epistemology. Hence for such Bayesians, Schupbach's account is not applicable to model-based RA.

the new model will give the same result, then there will be very little or no confirmation in this case. Clearly, then, we should prioritize eliminating the rival hypotheses that we think are plausible.[35]

In this section, I have argued that the application of Schupbach's account to model-based RA relies on several non-trivial assumptions. The first important assumption I discussed is that an agent *must* believe that the models they are considering are adequate representations of the target system not by mere luck but because, for whatever reasons, the models 'latch on' to the underlying mechanism in the target system. This is so because if it were not then, according to the agent, the target hypothesis would be an arbitrary conjunction (in other words, the hypothesis that the empirical hypothesis under investigation is correct and the hypothesis that the models are adequate for discerning whether or not it is correct would be unrelated to one another). Hence, without this assumption, the hypothesis that the empirical hypothesis under investigation is correct would be irrelevant to the explanation for the models' results and couldn't be confirmed by them. A second assumption I discussed is that logical hypotheses can be explanatory; I have argued that this is a necessary assumption for finding an adequate rival hypothesis, one that satisfies all Schupbach's conditions of ERA diversity. The final assumption I discussed is that the agent is not logically omniscient and that therefore, according to such an agent, the probability of the rival (logical) hypothesis can take non-maximal values.

Despite these substantial assumptions, I have argued that they may be reasonable in cases where the hypothesis we are interested in confirming through model-based RA is a 'robust theorem' (which I interpreted as the hypothesis that a causal structure of the model has a stable capacity to manifest a particular result). Hence, if this is right, Schupbach's account of ERA diversity can justify

[35]Indeed there might very well exist cases where we have good reasons to believe that a large set of the possible rival logical hypotheses are implausible. One of these reasons, for instance, could be that we *know* that all those models, despite involving different idealizations, 'belong to the same type, and they satisfy the Volterra principle, because they are of this type' (Raz 2017, 751). Indeed, as discussed in Section 3.2, Raz demonstrates that, as long as a condition that ensures that the average abundance of a system coincides with the relevant equilibrium is satisfied (see Raz 2017, 748), the Volterra principle holds for a more general model (a slight modification of one proposed by Gause (1934)). Hence, if an agent knows this, there won't be any confirmation when the models selected for RA belong to this type.

why and when model-based RA should increase one's confidence in a 'robust theorem'.[36] This is the good news.

There is some less good news, however. Firstly, under this account, the extent to which the target hypothesis H will be confirmed is highly contextual and will vary from individual to individual depending on their knowledge and beliefs about the derivation relationships in a family of models. Hence, even if two agents might agree that a set of models' results is ERA diverse with respect to H, they might strongly disagree about the extent to which this fact should confirm it. But of course, that's Bayesianism for you. Secondly, the above assumptions are not at all trivial assumptions and although in this section I have argued that they may (sometimes) be reasonable in cases where the hypothesis we are interested in confirming through model-based RA is a 'robust theorem', there is no reason to assume they are reasonable in other cases. Indeed, in Section 5.3, I will argue that in all cases where the hypothesis we want to confirm is that a result of a model is instantiated in the target system, and the models we select to check if that result is maintained involve incompatible assumptions about that target system, Schupbach's account of ERA diversity is inapplicable because not all of the above assumptions can reasonably hold.

[36] Of course, this only applies to agents who are Bayesian in the first place!

Chapter 5

The epistemic import of model agreement in climate science: what philosophers and scientists have to say about it

5.1 Introduction

The climate system is too complex to be faithfully represented so any attempt to model it so as to learn something about it will *necessarily* involve several idealizing and simplifying assumptions that scientists *know* fail to accurately represent the climate system. As Baumberger et al. remark, this is not in itself a problem since:

> The aim in climate modeling [...] is not (and cannot be) to arrive at a complete representation of the climate system that is correct in all details. The aim is rather to construct models that represent processes of the climate system in ways that make the model adequate for specific purposes.[1] (Baumberger et al. 2017, 4)

Let us then consider a case where all one cares about is whether a model is adequate for predictive purposes: couldn't one simply identify the model that is

[1] What it would mean for a model to be a 'complete representation of the climate system' is not at all a trivial matter in the first place.

most promising as a predictive tool? Looking at the models' histories of predictive successes and failures would be the most obvious place to start, but as Parker notes,

> [climate] models make predictions about what might happen 10 or 50 or 200 years from now under conditions that may or may not actually obtain during the intermediate years. [...] Weather forecasting models, by contrast, make predictions about what will actually happen over time periods of hours, days or weeks. Scientists can and do compile much information about the predictive strengths and weaknesses of these models. But for climate models, there is almost no such information, since the observational data that is needed in order to assess the quality of their predictions will not be available, even in principle, for quite some time. (Parker 2006, 353)

Could one perhaps look at simulations of past and present climate conditions and assess the model's predictive performance based on its *retrodictive* successes and failures? Unfortunately, as Parker further explains, this is also tricky:

> One serious problem [...] is that data are available for only a few quantities (e.g., temperature, pressure, precipitation), for only relatively recent time periods, and primarily for land locations and near-surface locations, and even these records are incomplete and of variable quality. Scientists lack a solid observational foundation against which to compare even the retrodictions of climate models. (Parker 2006, 353)

This is clearly bad news: if all climate models fail to be accurate representations of the climate system and there is not a straightforward method to test their predictive performance, there doesn't seem to be any good reason to rely on any single one of them for predictive purposes. But making no use of climate models until a single 'best' one can be identified is also hopeless, for two reasons. First, there is little reason to believe that it will *ever* be possible to identify a single 'best' model, for the same reasons as above. Second, we can't afford to wait. Hence as Parker remarks,

> Despite the fact that one must be careful when interpreting the results produced by multi-model ensembles, when it comes to addressing the global warming issue, the ensemble approach seems clearly better than the two most obvious alternatives, that is, relying on a single model and/or making no use of climate models until a single 'best' one can be identified. (Parker 2006, 361)

And indeed, the IPCC has certainly embraced an ensemble approach. As discussed in Section 2.4, the most recent Coupled Model Intercomparison Projection Phase 5 (CMIP5), for instance, was a huge collaborative effort, involving more than 20 climate modeling groups from around the world (Taylor et al. 2012, 486), to promote a standard set of model simulations whose outputs were then analysed by the AR5 authors to produce many of their findings.[2] Despite the harsh criticisms that I made in that chapter regarding the IPCC authors' *interpretation* of the results produced by multi-model ensembles, I have by no means questioned the IPCC ensemble approach itself.

But accepting that an ensemble approach is more reasonable than any other approach nonetheless gives rise to many questions: how should a model ensemble's results be interpreted? Should the fact that current climate models agree on a particular result (or a range of predictions) raise our confidence in that result? If so how much confidence? What kind of considerations are relevant to answer these questions?

This chapter is an exploration into the above questions. In Section 5.2, I will give a brief review of the 'exchange' between Lloyd and Parker on the epistemic import of robustness of current climate multi-model ensembles' results. I will conclude that none of these arguments that I consider in this section are able to shed light on the epistemic import of model robustness in climate science. In Section 5.3, I will turn to Winsberg (2018)'s claim that Schupbach's account of ERA diversity can finally help shed some light on the significance of the robustness of climate model ensembles' results. Unfortunately, we will see that there are strong reasons to be pessimistic here too. In Section 5.4, I will turn to the question of what climate scientists have said about the epistemic import

[2]The Coupled Model Intercomparison Project is now in its 6th phase.

of model agreement. In particular, I will focus on scientists' current (frenetic) search for an adequate measure of independence across climate models. After reviewing the various approaches that have been proposed to define and measure of the level of independence across models and the challenges that each of these approaches faces, I will argue that this search is implicitly guided by a undefended and questionable assumption: the assumption that the more dissimilar models are from other models in an ensemble, the greater the confidence we should have in the models' consensus.

5.2 Lloyd and Parker on the epistemic import of model robustness in climate science

As Winsberg (2018, 178) notes, Lloyd (2009, 2010, 2015) and Parker (2011) have been central figures in the debate about the epistemic import of model robustness in climate science. However, in contrast to Winsberg, it is in my view fair (and not just 'tempting')[3], 'to treat them as the robustness booster and robustness skeptic, respectively'. Hence, it is worth having a close look at what each has had to say about the epistemic import of model robustness in climate science, starting from Lloyd.

I find Lloyd's argument for the epistemic import of model robustness unclear in several respects. However, I will try to reconstruct her argument as best I can, in the hope that this will help me clarify exactly what is unclear about it and why I don't think it is a sound (nor valid) argument. To illustrate her argument for the epistemic import of model robustness in climate science, Lloyd (2015) considers the fact that 'all of the available climate models that incorporate greenhouse

[3]Winsberg 2018, 179) argues that in the exchange between Parker and Lloyd on robustness in climate science, they 'were addressing slightly different issues' since the targets of their discussions were somewhat different: whereas Parker focused on the question of 'how much work can RA do in understanding whether diverse models support climate hypothesis', Lloyd focused on the question of 'how much work can RA do in understanding how diverse models *and other sources of evidence* work together to support climate hypotheses'. Hence, according to Winsberg, it is somewhat a mistake to see Lloyd's and Parker's views in direct opposition to each other (see also Lusk's review of Winsberg's book for a similar take on this). However, contrary to Winsberg and Lusk, I do think that Lloyd and Parker did in fact address the very same question (i.e. the question of what is the epistemic import of model robustness in climate science, in light of the available evidence) and came to different conclusions as to the answer to this question.

gases as a cause of climate change produce an increase in global mean surface temperature (GMST) in the late 20th Century' as shown in the figure below.

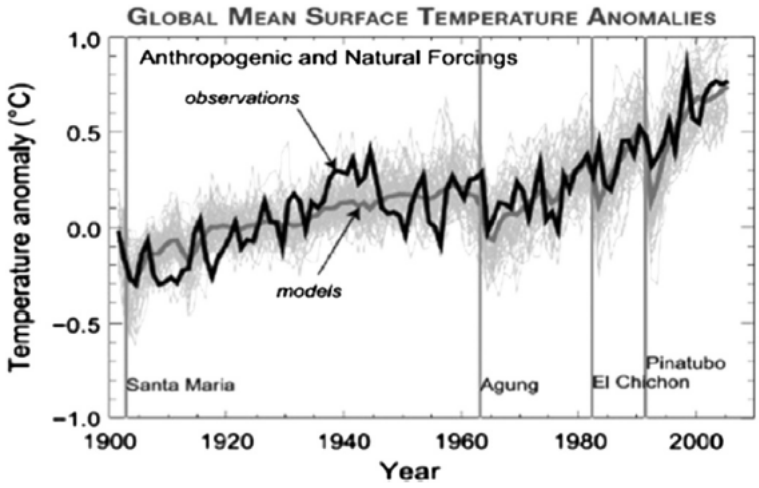

FIGURE 5.1: '14 GHG GCMs, in 58 simulations of 20th Century GMST anomaly trends..' (ibid., 61)

Lloyd argues that the fact that we have a robust and successful retrodiction from all these models is epistemically significant. I take her argument to consist of three steps. In the first step, Lloyd notes that although all these distinct models (M_i) differ in many of their assumptions and parameterizations, they all 'share a core representation of greenhouse gases (GHG) as a radiative cause' (ibid., 62). Reminiscent of Weisberg's analysis of model robustness, Lloyd then argues that we should think of this as the common causal core shared by these models, and hence this is a case in which we have 'a model-type M, which is characterized by the inclusion of the GHG causal core' and where 'there is variety of different assumptions and parameterizationsA_is [...] composing the rest of the model, such that (M&A_is) implies conclusion T' (ibid., 64), where T is the robust retrodiction in this case.

In the second step, Lloyd argues that each model in the ensemble is not only confirmed by its success in predicting T, as well as experimentation support for its GHG core causal process (e.g. Tyndall's and later laboratory experiments)[4],

[4]Already in the mid-19the century, a causal connection between CO_2 (and other greenhouse gases) and an increase in atmospheric temperature was supported by laboratory experiments done by John Tyndall (Hulme, 2009), but as Lloyd (2015, 62) remarks 'questions remained about whether these laboratory setups resembled the earth's real atmosphere enough to provide a causal explanation at the global scale'.

but also by a large set of observational evidence for its assumptions A_i, which will vary from model to model:

> Because of the variety of parameters, variables, and parameterizations used in the construction of $M_1 \ldots M_n$, there is also a wide variety of empirical evidence that can be brought to bear on the assumptions, A_is, of these individual models, in addition to scoring its empirical success in producing accurate global mean surface temperatures, T. For instance, one model may rely on empirical evidence supporting its parameter values in its modeling of the El Nino Southern Oscillation (ENSO), while another may rely heavily on the empirical support for a number of details, such as moisture content, drop size, etc., of its cloud parameterization. [...]Because the details of empirical support for these assumptions of the individual models the values relating to parameterizations, variables, parameter values, and model structures differ in the case of each individual model or model application, it is necessary to construct individualized sets of confirming empirical evidence for each model application in the set of robust models of the model family. Thus, the different A_is are each supported by their own bodies of empirical evidence, even while they produce competing or conflicting detailed climate systems. (Lloyd 2015, 63)

Before moving to the second step of Lloyd's argument, a clarification is important. Although Lloyd (2015) often talks of a *model* being confirmed by a variety of evidence, she clarifies that this is shorthand for the claim that a variety of evidence confirms the hypothesis that '[a model] represents specific aspects of the real world, say, various structures contributing to and predicting/retrodicting global mean temperature, to specified degrees, for purposes x, y, or z' (ibid., 64). So essentially what it means to confirm a model according to Lloyd is to confirm the hypothesis that a model is similar to a target in particular respects and degrees for the purposes of the modellers or those who use them. Indeed, as Parker

(2020) notes, the idea that models simpliciter are the objects of confirmation is a problematic one:

> On some accounts, scientific models are structures or objects; they are not the sort of thing that can be true or false and thus are not an appropriate target of confirmation. Even if a model is viewed as a complex hypothesis about the workings of a target system (per Oreskes et al. and many others), it is usually misguided to seek to confirm (or disconfirm or falsify) that hypothesis, since it is usually known from the outset to be false; some of the model's assumptions are known to be highly idealized or simplified, to appeal to fictional entities, and so on (Parker 2010). (Parker 2020, 458) [5]

However, as Parker remarks, although there is no sense in which one can confirm a model simpliciter there are a number of reasonable alternatives. Parker (2020) herself advocates an adequacy-for purpose view of model evaluation with which Lloyd's view seems to be compatible.

In the second step, Lloyd argues that since each model is independently empirically supported by a variety of empirical and experimental evidence, this makes each model 'a satisfactory candidate to serve as evidence or an "experiment."':

> We can imagine that each model is an "experiment" for purposes of a variety of evidence argument. These "experiments" are in the form of random, distinct, independently confirmed, models, $M_1, M_2, \ldots, M_n$, and their supporting observational and experimental evidence, in which GHG is part of the radiative causal core, and other assumptions, A_is, such as formulations of equations, values of forcings, or parameterizations, of the individual models vary. Significantly, each random model is well-supported by a variety of empirical and experimental evidence, making it a satisfactory candidate to serve as

[5] As Parker (2020, 458) further notes the idea that a model simpliciter can be confirmed is not merely an academic point since a view that sees models simpliciter as the object of confirmation can easily lead to misplaced confidence, such as confidence in any result obtained from the model even when this is not warranted.

> evidence or an "experiment." This situation includes, as we have discussed, that many of the A_is are often independently empirically supported, as well as the causal core itself having independent experimental and/or observational evidence of its own. (Lloyd 2015, 65)

From these two steps, Lloyd's concludes that:

> this is a way in which the GHG causal core itself can have its confidence and reliability raised through its repeated successes in producing accurate predictions/retrodictions of late 20th and early 21st C. global mean temperature, T, in conjunction with a variety of independently empirically supported model assumptions. Model robustness describes a pattern of models and evidence, which is described within a variety-of-evidence inference, as telling us more than any given piece or subset of pieces of evidence as used in these inferences, and as giving us increased confidence first in the causal core, and ultimately in the model outcomes. (Lloyd 2015, 65)

I find Lloyd's argument confusing for at least two reasons. The first is that it is not sufficiently clear what hypothesis model robustness is supposed to confirm according to Lloyd. In the quote above Lloyd argues that model robustness in this case should increase our confidence in the *causal core* of the models, which I take to be the hypothesis that there is a causal connection between greenhouse gases and increases in atmospheric temperature (but not the hypothesis that greenhouse gases necessarily increase temperatures in the earth's real atmosphere). However, she also argues that model robustness in this case will 'ultimately' also increase our confidence 'in the model outcomes' (ibid., 65). But what outcomes is Lloyd referring to is not clear. The case that she considers to illustrate her argument is one in which we have a robust and *successful retrodiction*, so it is hard to see why we would want to confirm this outcome in the first place. Perhaps then she is thinking of *other* various predictions by the model ensemble in question? But what other predictions she may be referring to is not clear. In yet other passages she argues that model robustness raises 'the

confidence connecting the causal core, GHG, of the model-type, *M*, to the 20th and early 21st Century warming outcomes, to specified degrees and respects, and assuming a particular purpose' (ibid., 65) or that it increases 'confidence in that causal core as a good explanation of the robust and verified model predictions/retrodictions' (ibid., 67). So here she seems to argue that model robustness can increase our confidence that greenhouse gases have played a substantial causal role in the 20th and early 21st century temperature increase. These are all *different* hypotheses, and without clarity as to *which* one model robustness is supposed to confirm in this case it is hard to understand where the epistemic import of model robustness actually lies according to Lloyd.

The second reason for why I find Lloyd's argument confusing (or better: puzzling) is that it is not at all clear what to make of the idea that each model in the ensemble is a 'satisfactory candidate to serve as evidence or an "experiment". That is, what can it possibly mean to take a *model* as evidence for a hypothesis? For instance, if one understands a model as a complex hypothesis about the global climate, then it doesn't seem conceptually coherent to think of this complex hypothesis as evidence for a hypothesis, nor if we think of models as structures or objects for that matter. So whatever Lloyd means by the claim that a model can serve as evidence must mean something else, but Lloyd herself doesn't clarify this.

I am clearly not alone in this puzzlement. Indeed, in his attempt to reconstruct Lloyd's variety of evidence argument for the epistemic import of model robustness Justus (2012) is confronted with the very same puzzlement. In his attempt to reconstruct Lloyd's argument, Justus (2012) applies Fitelson (2001)'s account of confirmational independence (discussed in Section 4.2.3) as follows:[6]

> Because each [Global Climate Model] GCM is confirmed by various predictions, perhaps they can be treated as bits of evidence for the common core *C* that they share. Returning to Fitelson's account and making the relevant substitutions, the generalization would require what follows:

[6] Although, Lloyd (2015) does not mention Fitelson's account of confirmational independence (nor any other account of confirmation for that matter) to justify her variety of evidence argument in this instance, she does in her earlier papers; see Lloyd (2009, 2010).

> If GCM_1 and GCM_2 individually confirm C and are [confirmationally independent] regarding the (core) hypothesis C, then $c(C, GCM_2 \& GCM_1) > c(C, GCM_1)$, and $c(C, GCM_2 \& GCM_1) > c(C, GCM_2)$.[7]
>
> But this is flawed on many fronts. First, since C is part of GCM_i, the right side of each [inequality] seems to be 0, and the first part of the preceding antecedent, false: GCM_i deductively entails C, but that certainly does not establish that it confirms C. And, second, since GCM_i and GCM_j ($i \neq j$) are logically incompatible hypotheses about global climate, the left-hand side of each [inequality] seems undefined: the conditionalizations are predicated on an impossible circumstance. (Justus 2012, 805)

Let's focus on the second flaw that Justus points out with this analysis, the fact that 'since GCM_i and GCM_j ($i \neq j$) are logically incompatible hypotheses about global climate, the left-hand side of each [inequality] seems undefined: the conditionalizations are predicated on an impossible circumstance.' Indeed, if we take *models* to serve as evidence for a hypothesis, as suggested by Lloyd, and if we understand models to be complex incompatible hypotheses about the global climate, then in order to apply Fitelson's account in this instance we would have to assume that a set of incompatible hypotheses about the global climate can confirm a hypothesis. But this can't be right since if the probability of the 'evidence' is 0, conditionalization is undefined and hence under any plausible confirmation measure, the confirmatory value of learning this evidence will also be undefined. What I (and arguably Justus too) take this to show is that Lloyd's idea that *models* can serve as evidence for a hypothesis is conceptually problematic.

There is, however, a way to improve (if not save) Lloyd's argument. Rather than treating *models* as evidence for a hypothesis H, we might be able to treat their *results* as evidence for a hypothesis. So returning to Fitelson's account and making the relevant substitutions, the generalization would now require what follows:

[7] In the original quote those inequalities are equalities. But I changed them, as it is a typo (Jack Justus has confirmed this to me).

> If R_1 and R_2 (which are the results of GCM_1 and GCM_2 respectively) individually confirm a hypothesis H and are confirmationally independent regarding H, then $c(H, R_2\&R_1) > c(H, R_1)$, and $c(H, R_2\&R_1) > c(H, R_2)$.

In contrast to Justus's application of Fitelson's account, this is one doesn't seem to be conceptually incoherent.[8] However, despite this, there are in fact plenty of reasons to doubt that it is reasonable to assume that the models' results R_1 and R_2 are confirmationally independent regarding H. For even if we were to accept Lloyd's first step - that distinct 'models' are supported by distinct bodies of evidence - it is really unclear why this alone should convince us that this assumption is reasonable. As discussed in Section 4.2.3, if the models in question share idealizations, uncertain assumptions, omissions etc. it is unreasonable to assume their results to be confirmationally independent regarding a hypothesis. And as Parker (2011, 591) notes, this is indeed the case for current climate models since 'there are climate system features and processes [...] that are not represented in any of today's models but that may significantly shape the extent of future climate change on space and time scales of interest. In addition, when it comes to features and processes that are represented, different models sometimes make use of similar idealizations and simplifications.'

Hence, even if a more charitable reconstruction of Lloyd's variety of evidence argument is possible, there is very little reason to think that the assumptions on which this reconstruction relies are satisfied by current climate model ensembles (and future climate model ensembles too for that matter).[9] Perhaps Lloyd might not want to rely on Fitelson's account of confirmational independence after all.

[8] By this I don't mean to suggest that Justus wasn't well aware of the conceptual incoherence involved nor that Justus was making some kind of error in attempting to reconstruct what possibly Lloyd could have had in mind. Indeed, it is clear that the aim of Justus's analysis above was to reveal the incoherence of Lloyd's approach.

[9] My sceptical attitude towards Lloyd's argument for the epistemic import of model robustness is not *at all* meant to suggest that I don't think there is overwhelming evidence supporting the hypothesis that greenhouse gases are responsible for the 20th and early 21st century temperature increase. Indeed there are plenty of established results (justified independently of GCMs) that provide overwhelming strong evidence in support of it (e.g. our very good understanding of the causal mechanism of the greenhouse effect, our identification of water vapour, carbon dioxide as greenhouse gases, Paleo climate data revealing that current CO2 concentration levels far exceed natural fluctuations etc.). However, the fact that we have overwhelming evidence in support of this hypothesis has nothing to do with whether or not Lloyd's argument for the epistemic import of model robustness is a good one.

However, given that she does not provide any other account of confirmation, her argument is at best incomplete.

Parker (2011), in contrast to Lloyd, has a rather more pessimistic outlook on the epistemic import of model robustness in climate science. She considers several arguments that could in principle be used to justify why we should have high confidence in current climate ensembles' robust predictions, but she concludes that all those arguments rely on very questionable assumptions and that they therefore not 'readily applicable in the context of ensemble climate prediction today'. Here I will only mention two of the arguments she considers (I only focus on the ones that don't rely on any sort of assumption of probabilistic independence since as already discussed extensively in Chapter 4, I don't think the epistemic import of model robustness can be adequately defended on the basis of any sort of probabilistic independence). One argument she considers is the following (ibid., 584):

1. It is likely that at least one simulation in this collection is indicating correctly regarding hypothesis H.

2. Each of the simulations in this collection indicates the truth of H.

$\therefore$ It is likely that H.

She argues that there are at least two possible approaches to justify why the likely adequacy condition (premise 1) is met by today's multi-model ensembles: one that focus on ensemble construction and one that focuses on ensemble performance. Under the first approach: 'one would argue that an ensemble of models samples so much of current scientific uncertainty about how to represent the climate system (for purposes of the predictive task at hand) that it is likely that at least one simulation produced in the study is indicating correctly regarding H' (Parker 2011, 584). However, as already mentioned above, as far as today's multi-model ensembles are concerned this approach is unlikely to succeed since 'these ensembles are ensembles of opportunity [...] they are not designed to span an uncertainty range' (ibid., 585). On the performance approach, on the other hand: 'an ensemble is viewed as a tool for indicating the truth/falsity of

hypotheses of a particular sort, of which the predictive hypothesis H is an instance; the ensemble's past reliability with respect to H-type hypotheses is cited as evidence that it is likely that at least one of its simulations is indicating correctly regarding this particular H' (Parker 2011, 585). The main problem Parker raises with using this approach to justify the likely adequacy condition (premise 1) has to do with the tuning of climate models: 'given the ad hoc nature of the tuning process, and the fact that today's climate models are far from perfect in their representation of the climate system, it cannot be assumed that the performance of a tuned climate model with respect to as-yet-unseen data will be similar to its performance with respect to the data to which it is tuned. Moreover, when today's climate models are tuned, it is often difficult to adequately test their out-of-sample performance, both because reliable observations of past climate are limited and because most observations that are available are for time periods in which greenhouse gas concentrations were significantly lower than they are expected to be in the future.' (Parker 2011, 587)

Parker (2011, 590) also considers the following Bayesian argument for why agreement across models should increase confidence in the common result:

1. e warrants significantly increased confidence in predictive hypothesis H if $p(e|H) >> p(e|\neg H)$.[10]

2. $e =$ all of the models in this ensemble indicate H to be true.

3. The observed agreement among models is substantially more probable if H is true than if H is false; that is, $p(e|H) >> p(e|\neg H)$.

$\therefore$ e warrants significantly increased confidence in H

However, she strongly doubts that the third premise can be adequately justified as far as today's climate model ensembles are concerned. In particular she argues that there are many reasons to worry that climate models might all indicate the truth of a predictive hypothesis, despite it being false:

[10]This premise follows from Bayes' theorem.

> First, there are climate system features and processes— some recognized and perhaps some not—that are not represented in any of today's models but that may significantly shape the extent of future climate change on space and time scales of interest. In addition, when it comes to features and processes that are represented, different models sometimes make use of similar idealizations and simplifications. Finally, errors in simulations of past climate produced by today's models have already been found to display some significant correlation (see, e.g., Knutti et al. 2010; Pennell and Reichler 2011). Thus, in general, the possibility should be taken seriously that a given instance of robustness in ensemble climate prediction is, as Nancy Cartwright once put it, "an artifact of the kind of assumptions we are in the habit of employing" (1991, 154). Perhaps with additional reflection and analysis, persuasive arguments for $p(e|H) >> p(e|\neg H)$ can be developed in some cases, but at present such arguments are not readily available. (Parker 2011, 591)

This ends my brief review of the 'exchange' between Lloyd and Parker as far as the epistemic import of *model* robustness in climate science is concerned. As we have seen, none of the arguments that I have considered in this section (one from Lloyd and two from Parker) seem to provide a satisfactory justification for why robust predictions/retrodictions should increase (let alone substantially increase) our confidence in climate hypotheses. However, Winsberg (2018) has recently suggested that Schupbach's account of ERA diversity can help shed some light on the significance of the robustness of climate model ensembles' results. The aim of next section is to critically assess whether this is in fact the case.

5.3 Winsberg on ERA diversity and Climate model ensembles

In his recent book 'Philosophy and Climate Science', Winsberg (2018) has emphatically argued that Schupbach's account of ERA diversity can finally shed

light on the epistemic import of model robustness in climate science.[11] According to Winsberg:

> We should not ask whether the outputs that are robust under the ensemble of opportunity of models should be trusted. This question is too simple to have a determinate answer. Rather, we should talk about specific climate hypotheses, and inquire about how [ERA]-diverse our ensemble of models is with respect to each one of these hypotheses individually. (Winsberg 2018, 193-94)

From this view it follows that:

> Whether or not an ensemble of models is a good candidate for lending strong support for a hypothesis via RA depends almost entirely on the extent to which the set of models suffices for ruling out competing hypotheses. This means that just because the set of procedures we have that detect H are ERA-diverse does not imply that we should have confidence in H. [ERA]-diversity only implies CEP [cumulative epistemic power], i.e. it only implies that you are headed down the road to acceptance as you increase the size of the set of procedures. Once we know that a set is [ERA]-diverse the question of whether it is large enough to warrant acceptance of H, whether it is sufficiently [ERA]-diverse, is a further question. And the answer to that further question will always be a matter of judgment, context, considerations of inductive risk, etc. (Winsberg 2018, 194)

In Section 4.3, I argued that the application of Schupbach's account to model-based RA relies on several non-trivial assumptions. Despite this, I argued those assumptions may be reasonable in cases where the hypothesis we are interested in confirming through model-based RA is a 'robust theorem' (interpreted as the hypothesis that a causal structure of the model has a stable capacity to manifest

[11] And he is not alone. According to O'Loughlin (2021, 36), 'Winsberg (2018) convincingly argues that [Schupbach's account] can be applied to climate models.' In reviews of Winsberg's book, Lusk (2019) writes that 'Winsberg's argument is a convincing reconceptualization of robustness analysis in climate science' and Knüsel (2020, 116) that 'Winsberg [...] makes a novel, convincing suggestion for when multiple sources of evidence in favor of a hypothesis are meaningful in climate science.'

a particular result). In this section, however, I will argue that Schupbach's account is inapplicable to a very large class of model-based RA, that is, all cases in which the hypothesis we want to confirm is that a result of the model is instantiated in the target system and the models we select for RA involve incompatible assumptions about that system. Hence, contrary to what Winsberg suggests above, I will conclude that Schupbach's account is not applicable in the context of climate model ensembles. Hence, we should not rely on this account to help us shed some light on the significance of the robustness of climate model ensembles' results. But first, it will be useful to look at one of the main sources of uncertainty in climate modeling: the parameterization of physical processes.

Several physical processes, whose representation is thought to be critical in generating accurate projections, cannot be resolved directly by current climate models since they occur at a smaller scale than the models' grid resolution.[12] For instance, the development and evolution of cloud processes are thought to play a very important role in the Earth's radiation budget. However, these processes cannot be resolved directly by current climate models since clouds can be as diminutive as a few hundred metres across – substantially smaller than the current models' grid resolution (around 50–100 km horizontally) (Parker, 2013). Therefore, in order to include the effects of these subgrid processes in the evolution of the model variables, these subgrid processes must be represented in terms of larger-scale variables. This process of representing physical processes that cannot be resolved directly by the model is known as parameterization. Since a parameterized subgrid process is one for which the model has no direct information, the subgrid process must be related to known model variables in one way or another. Hence the process of parameterization involves both a choice of what equations should describe the various relationships between the subgrid process and the known model variables and a choice of the various

[12]Climate models represent the atmosphere by a three-dimensional set of points (called a grid). So model resolution (which depends on the number of discrete grid points) refers to the horizontal and vertical scales that can be resolved by the model. Clearly, the higher the resolution, the more physical processes can be directly resolved. However, there will always be some physical processes that cannot be explicitly represented (and hence will have to be parameterized) *regardless* of the model's resolution: either because they occur at too small a scale (e.g. the formation of cloud droplets occurs on the molecular scale) or because of their complexity (e.g. biochemical processes of vegetation).

parameter values within those equations. It is clear, then, that when it comes to the parameterization of subgrid processes, there will always be at least two sources of uncertainty. One is *parameter uncertainty*, i.e. uncertainty concerning the adequate parameter values; the other is *structural uncertainty*, i.e. uncertainty concerning the adequate equations describing the relationships between the subgrid process and known model variables.[13] However, it is important to note that regardless of these choices, parameterizations 'by necessity distill only the essential aspects of the physical processes they represent' (Stensrud 2007, 9); in other words, *all* parameterizations are invariably simplified and idealized representations of complex physical processes.

With this in mind, let us now look at an example on which Winsberg relies to convince us that Schupbach's account of ERA diversity can shed light on the epistemic import of model-based RA in climate science:

> Suppose that a climate simulation can be used to calculate that equilibrium climate sensitivity (ECS) is greater than 2°C. One explanation of this is that ECS is actually greater than 2°C. Thus this would count as a detection of the hypothesis (that ECS is greater than 2°C) by a model. But another possible explanation might be that the calculated result is an artifact of the large grid size of the simulation. A natural move is to try to halve the grid size and check to see if the result is maintained. If it is half the grid size again. If the result remains stable, then the probability of that rival explanation goes way down. Thus a reasonable ensemble of different simulation models with descending grid size could count as [ERA] diverse. [...] But even once we are convinced that the grid size is not responsible for the purported detection of the hypothesis, there remains the possibility that the detection is an artifact of the way that cloud formation is

[13] An important difference between structural and parameter assumptions is that for any given structural assumption, the space of possible alternative structural assumptions is, arguably, undefined - since there is no clear way to circumscribe the class of possible alternative equations describing the various relationships between the subgrid process and the known model variables. In contrast, for a given parameter assumption, the space of possible alternative parameter assumptions is well defined: it is the space of possible numerical values (but as Baumberger et al. (2017, 10) remark 'it is nonetheless computationally intractable to its dimensionality').

> parameterized in the simulation. A rival cloud parameterization can be tried. *Certainly those two methods of detection would count as ERA diverse*. Again, context and judgment, but this time presumably of a more subtle and difficult character, would be required to decide at what point, if any, enough different cloud parameterization schemes are enough to rule out all such hypotheses. (Winsberg 2018, 192-93, my emphasis)

Winsberg actually uses two examples to illustrate when simulations' results count as ERA diverse. The first concerns simulations with different grid sizes, the second concerns simulations with different parameterizations. However, the first example is, in my view, besides the point when it comes to discussions concerning the epistemic import of model-based RA in climate science. Scientists are very well aware that higher resolution would, for instance, reduce the influence of physical parameterizations of some of the processes that are sufficiently well understood, but that occur at finer spatial and temporal scales. So if they could increase resolution, they would! But substantially increasing model resolution for global climate models is not at all trivial and requires an enormous amount of computing power. Hence for the time being, the resolution of current global climate models is what it is and climate scientists have to live with it. The question that matters to us is whether *current* global climate models can be used to learn about the climate. In particular, in this case, what we want to know is whether the fact that a result is robust across current multi-model ensembles should increase our confidence in that result. Given that those models often include distinct parameterizations for the same physical process, Winsberg's second example seems more pertinent to this question.

In this example, Winsberg considers a simulation involving a particular parameterization scheme for cloud formation (i.e. a particular structural assumption B_1) which gives result R_1. He then claims that if one were to observe that a second simulation involving a rival cloud parameterization scheme (i.e. a different structural assumption B_2) gives the same result R_2, these two detections would count as ERA diverse. Hence, Winsberg is implicitly assuming that it

is possible to find a target explanation H and rival explanation H' that satisfy Schupbach's conditions of ERA diversity in this case. The following candidates for H and H' may, prima facie, seem reasonable:

> H: ECS is greater than 2°C (R) & both climate simulations are adequate (not by mere luck) representations of the target system,
>
> H': The original climate simulation entails R_1 & if B_1 is replaced with B_2 the new simulations entails $\neg R_2$.

However, I will argue that, due to the incompatibility of the assumptions B_1 and B_2, it must be the case that either H or H' fails to be a plausible candidate. Hence, contrary to what Winsberg claims, I will conclude that 'certainly those two methods of detection would [*not*] count as ERA diverse'.

In a nutshell, my argument is the following. In light of the incompatibility between B_1 and B_2, there are only two possible epistemic states for an agent to be in, and under neither of them is it possible for an agent to find both an adequate target and rival hypothesis that satisfy all of Schupbach's conditions of ERA diversity. They are the following: 1) an agent believes that at most one of these two simulations can be adequate (not by mere luck) for the purpose at hand and hence H is not a plausible candidate for the target hypothesis since for such an agent $Pr(H) = 0$; 2) an agent believes that both simulations can be adequate (again not by mere luck) and hence H' is not an adequate rival hypothesis since for such an agent $Pr(H') = 0$.

Consider case 1). An agent may reasonably believe that since different parameterizations for a particular process (in this case, cloud formation) are *competing* ways to represent such a process (Parker, 2006), then at most one of these two simulations can be adequate (not by mere luck) for the purpose at hand. Hence, for an agent in this epistemic state $Pr(H) = 0$, and hence H is not a plausible target hypothesis for them. Consider case 2). An agent may reasonably believe that although different parameterizations for a process are competing ways to represent such a process, those differences are *irrelevant* for whether or not the simulations are adequate for the purpose at hand. According to such an agent, since the simulations are sufficiently similar in what they consider to

be *all* the relevant aspects, it is possible to assume that *both* simulations can be adequate (not by mere luck) for the purpose at hand. Hence, for such an agent H is a plausible target hypothesis. However, for an agent to believe that those differences are irrelevant for whether or not the simulations are adequate (not by mere luck) for the purpose at hand, they must effectively believe that the differences across those simulations are irrelevant to the result they will produce. In other words, such an agent must believe that both simulations, despite their differences, are bound to give the same result. Hence according to such an agent, H' is false and is thus not a plausible rival hypothesis.

A more general perspective might further help us to see what is at stake here. I take the above example to be an instance of a more general class of model-based RA, one in which in light of the modelers' uncertainty about how to adequately represent a target system, there are many possible incompatible ways to do so and, perhaps surprisingly,[14] it is discovered that all the models available agree on a particular result. For Schupbach's account of ERA diversity to apply, and hence reveal why this fact should increase one's confidence that this result is instantiated in the target system, it must be possible to provide a partly empirical explanation for this coincidence, one according to which the *truth* of such a result plays not only a necessary role but also a *sufficient* one (aside from representational considerations). I argue that although an explanation according to which the truth of such a result plays a necessary role is possible, one according to which it plays a sufficient one is not. This is because the truth of the result cannot on its own explain why all those models *agree* on that result. In order for the truth of the result to explain why those models agree on that result, we must also independently believe that the differences across the models are irrelevant to the result they will give. In other words, the truth of the result can perhaps explain why the models agree on *that* result, but not why they *agree* in the first place. Hence, if this is right, Schupbach's account is not applicable to these cases, because for the truth to explain this coincidence, one must presuppose

[14]Perhaps not. Indeed this very much depends on how we understand "agreement". For instance, there is nothing surprising about the fact that incompatible models may all agree that the value for a particular variable is within a given range, especially if that range is determined after observing the models' results, which is how Winsberg himself seems to understand agreement.

the models, despite their differences, are bound to give the same result (since those differences are irrelevant for the purpose at hand). Hence, by accepting that the truth of a result can be part of an adequate explanation for why all these models give that result, one must at the same accept that a rival explanation that satisfies Schupbach's conditions of ERA diversity is not possible in this case.

There is a possible objection to my argument that is glaring and that I should respond to before concluding. My argument relies on the idea that if an investigator considers the hypothesis H to be plausible, then she must assign zero probability to the incompatible H'. But surely, one might object, a Bayesian investigator could assign both H and the incompatible H' some probability > 0; she can be uncertain about which is true. She collects more evidence precisely because she wants to discriminate between them. However, here is why I don't think this objection works. This objection relies on the idea that the investigator is uncertain about whether or not the differences across the models are irrelevant to the result they will produce. That is, according to the investigator, the models she considers might or might not give the same result, she simply is unsure. But a precondition for H to be true (and hence for all the models in the ensemble to be adequate representations of the target system despite making incompatible assumptions about the target system) is that all the models in the ensemble will necessarily give the same result whether or not it holds in the target system. Since the investigator does not know whether this is the case, this hypothesis must now be part of the target hypothesis H. Hence the target hypothesis H must state something like the following:

> H: R holds in the target system & both climate simulations will necessarily give the same result whether or not it holds in the target system & both climate simulations are adequate (not by mere luck) representations of the target system

But notice that the above hypothesis can be equivalently rewritten as follows:

> H: R holds in the target system & the first simulation is an adequate (not by mere luck) representation of the target system & so is the

> other simulation because it is bound to give the same result as the first one whether or not it holds in the target system

Which can further be equivalently rewritten as follows:

> H: R holds in the target system & the first simulation is an adequate (not by mere luck) representation of the target system and hence it gives result R & if the first simulation gives result R so must the second simulation whether or not R holds in the target system.

But then notice that under H, the fact that the second simulation gives result R has nothing to do with whether or not R actually holds in the target system. Under H the second simulation gives R merely because it is bound to give the same result as the first simulation independently of whether R holds in the target system. Hence H in this case is really an arbitrary conjunction of two hypotheses H_1 and H_2 where H_1 is the hypothesis that R holds in the target system and the first simulation is an adequate representation of the target system and H_2 is the hypothesis that the second simulation must give R independently of whether R holds in the target system since the first simulation gives R. But then H_1 is clearly irrelevant to the explanation of why the second simulation gives result R and hence cannot be confirmed by it. Hence, given that the ultimate aim is to confirm that R holds in the target system, H cannot be an adequate target hypothesis.

In this section, I have argued that when the hypothesis we want to confirm is that a result of a model is instantiated in the target system, and the models we select to check if that result is maintained involve incompatible assumptions about that target system, Schupbach's account of ERA diversity is inapplicable, for it is impossible to find both an adequate target and rival hypothesis that satisfy all conditions of ERA diversity.

Indeed, I think the idea that there must be an explanation for the robustness of a result in order for this robustness to be epistemically significant is simply the wrong way to think about what is going on in these cases of RA. What matters in these cases is not *why* the models agree on a result, but rather that they do at all!

For instance, as discussed above, there is a great deal of uncertainty about how to adequately represent the climate system; there are thus many ways one might attempt to do so. The hope, then, is that *at least* one of the available models is adequate for the purpose at hand, not necessarily all of them![15] Therefore, one way to motivate the idea that our confidence should increase the more models agree on a result is by arguing that by considering those additional models we can increase our confidence that at least one of the selected models is adequate for the purpose at hand. This is a very different way of motivating the epistemic import of model-based RA, and notice further that it has very little, if anything, to do with the agent's knowledge and beliefs about the derivational relationships in a family of models. Unfortunately, however, although I think this is the right way to think about what's going on in these cases of model-based RA, this argument doesn't take us very far without an understanding of what is the space of (possibly) adequate representations of the target system in question, and the extent to which the models we select are relevant for spanning that space. These are, in many cases, very hard questions, questions that we philosophers *should* help with if we want to help provide an adequate justification for the epistemic import of model robustness in those cases.

Indeed, I think there is a lot for us philosophers to think about in relation to those questions. As we will see in the next section, climate scientists have been trying for some time to find a measure of independence that can satisfactorily capture how dissimilar climate models are from one another (e.g. Bishop and Abramowitz, 2012; Sanderson et al., 2015; Annan and Hargreaves, 2017; Boe, 2018). Although these attempts vary considerably, the implicit assumption motivating all of them is that the more dissimilar models are from other models in an ensemble, the greater the confidence we should have in the models' consensus. But why *should* we assume that the more dissimilar an ensemble is, the more it spans current scientific uncertainty? For instance, two models may be rather similar in most respects and yet involve different parameterization schemes for highly uncertain processes (e.g. cloud formation). Although we might judge

[15] But as discussed in Section 5.2, as far as today's multi-model ensembles are concerned, this can at best be a hope rather than a justified assumption, for 'these ensembles are ensembles of opportunity [...] they are not designed to span an uncertainty range' (Parker 2011, 585).

these models to be rather similar overall, they might span more scientific uncertainty about how to adequately represent the climate system than two models we might judge less similar overall but that do not involve different parameterization schemes for such highly uncertain processes. That is, considerations on dissimilarity across models do not on their own seem to be sufficient for assessing the extent to which an ensemble samples current scientific uncertainty. So what *are* the relevant considerations? And which considerations can actually be implemented in practice? I believe *these* are the kinds of questions that we, philosophers, should think about if we are genuinely interested in helping scientists evaluate the epistemic import of model-robustness in climate science. Hence, it is time to turn to those.

5.4 Independence revisited: what have climate scientists said about the epistemic import of model agreement?

In a rough survey of the contents of several leading climate journals, Pirtle et al. (2010) found 188 articles 'in which the authors relied on the concept of agreement between models to inspire confidence in their results' (353). Indeed, it is not hard to find quotes by climate scientists in which they make explicit the thought that if multiple models agree on a result this should increase our confidence that that result holds in the actual world. For instance, Lambert and Boer (2001, 88) write that 'A small value of δ indicates agreement among models and supports the assumption that they are capturing the processes that govern that variable and hence its climate. A large value of intermodal scatter, on the other hand, indicated disagreement and unreliability'. Tebaldi et al. (2011, 1) write 'if multiple models, based on different but plausible assumptions, simplifications and parameterizations, agree on a result, we have higher confidence than if the result is based on a single model, or if models disagree on the result'. Boe (2018, 2771) write that 'At the core of the multi-model approach lies the basic idea that if the results of an additional model B are close to the ones of a model A, then our confidence in the results of A is reinforced. Obviously, it is only true insofar A and B are not near identical'.

But not everyone thinks that model agreement is always confirmatory, or at the very least not everyone thinks that all cases model agreement are equally confirmatory. Indeed a prominent idea among climate scientists is that *independence* across models is crucial for their agreement to be confirmatory (or perhaps substantially confirmatory). However, the notion of independence that they have in mind is not always clear and can be used to mean very different things. Indeed, there are at least three different interpretations of independence that have been discussed in the climate literature and the distinction between them is significant in many respects. Broadly they can be characterized as follows:

1. Under the first interpretation, 'the assumption of independence is equivalent to the interpretation that each model approximates the real world with some random error' (Knutti et al., 2010). This is often referred to as the "truth plus error" hypothesis/paradigm;

2. Under the second interpretation, the degree of independence is determined by the amount of divergence of models' outputs independent of observations (Abramowitz and Gupta, 2008) or by the degree of correlation of observed model errors (Bishop and Abramowitz, 2012; Sanderson et al., 2015). Under this interpretation, independence is measured *a posteriori*;

3. Under the third interpretation, the degree of independence is determined by the degree of shared formulation in the models. Hence under this conception of independence models are classified 'based on the independence of their structure' (Abramowitz, 2010). Under this interpretation, independence is measured *a priori*.

Notice that under the first interpretation, independence is not a matter of degrees. In other words, under the first interpretation models are either independent or they are not. Under the second and third interpretation, on the other hand, independence is a matter of degrees. In other words, under these interpretations models can be more or less independent and what we are interested

in is the extent to which the are. Below I will discuss the first interpretation, and in the next subsection, I will turn to the second and third interpretations.

Under the first interpretation of independence what it means for models to be independent is that their errors are independent and identically distributed (typically assumed to be normally distributed with zero mean). This is referred to as the "truth plus error" hypothesis/paradigm. And as Knutti et al. (2010, 2745) remark many Bayesian methods that are used to interpret the results derived from multi-model ensembles rely on the assumption that the truth plus error hypothesis is true and according to Leduc et al. (2016, 8302) 'the truth-plus-error paradigm remains the most widely used technique for processing multimodel ensemble'.[16]

As Annan and Hargreaves (2017) point out, if the truth plus error hypothesis were actually true it would have rather remarkable consequences:

> Although it has not generally been explicitly stated, even a small ensemble of samples drawn from such a distribution would be an incredibly powerful tool. If we could sample models from such a distribution, then we could generate arbitrarily precise statements about the climate, including future climate changes, merely by proceeding with the model-building process indefinitely and taking the ensemble mean. This would obviate the need both for computational advances and also for any additional understanding of how to best simulate the climate system [...] For example, if we accept the arguments of Pennell and Reichler (2011) that the CMIP3 ensemble contains eight "effectively independent" models then its full range of sensitivity values, 2.1–4.4 C, would still be a legitimate 99% confidence interval for the true sensitivity [...] The same argument would apply to any other output or derived parameter of the model climates.[17] (Annan and Hargreaves 2017, 212-13)

[16] Indeed as discussed in chapter 2 (section 2.5), the IPCC seems to often implicitly rely on this assumption too (but we have also seen that it is not very clear on what other assumptions they are relying to interpret climate model ensembles results as they do).

[17] Annan and Hargeaves (ibid., 213) further remark that this would imply that 'we could be "virtually certain" (to use the IPCC calibrated language) that the model ensemble bounds multiple aspects of the behaviour of the climate system, even with this very modest number of number of

As a further example of the impact of the error plus truth hypothesis, consider figure 5.2 below taken from Knutti et al. (2010), showing various pdfs obtained using a Bayesian method developed by Furrer et al. (2007) which relies on the truth plus error assumption. As the number of models increases the uncertainty in the true value of the temperature change (i.e. the width of the probability density function) decreases substantially as the number of models included in the ensemble increase from 4 to 21 models.

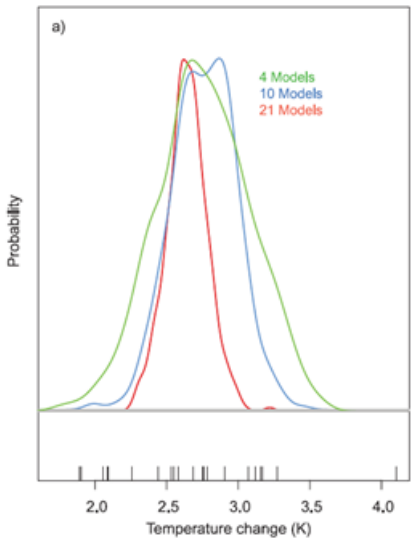

FIGURE 5.2: 'PDFs for annual global temperature change for the period 2080–99 relative to period 1980–99 from the Bayesian method by Furrer et al. (2007), for the A1B scenario and for 4, 10, and 21 models.' (Knutti et al. 2010, 2746)

Notice further that by applying Furrer et al.'s Bayesian method, we can conclude that we should be "virtually certain" (to use the IPCC calibrated language) that the value for the temperature change lies within the range of the values predicted by the 21 models. This is anything but a humble conclusion. [18]

But however striking the consequences of the truth plus error assumption, and despite the fact that many Bayesian methods that are used to interpret the results derived from multi-model ensembles rely on it, there is in fact very little reason to think it is plausible in most if not all cases. Indeed there are many

"effectively independent" models'. However, it is worth noting that this is a (alas very common) fallacy, one that Morey et al. (2016, 104) call *The Fundamental Confidence Fallacy*: 'If the probability that a random interval contains the true value is X%, then the plausibility or probability that a particular observed interval contains the true value is also X%; or, alternatively, we can have X% confidence that the observed interval contains the true value.'

[18] And (strikingly) one that was already obtained by considering only 4 models!

studies that show that models' errors are often correlated and that the mean of an ensemble does not converge to the truth as the number of models in an ensemble increases as should be the expected if the error plus truth hypothesis were true. Knutti et al. (2010), for instance, show that the errors of the models' results in the CMIP3 are strongly correlated and that mean of the CMIP3 does not asymptotically converge to observations. But crucially, the knowledge that climate models often share many simplifications, limitations and assumptions should already provide enough of a reason to suspect that this assumption is not appropriate in the first place as many have noted (e.g. Knutti et al. 2010, Bishop and Abramowitz 2012, 4)).[19]

In light of the implausibility of the truth plus error assumption, climate scientists have been frenetically trying to find other approaches to define (and measure) independence across models. To the best of my knowledge, all of these alternative approaches either rely on the second interpretation of independence or on the third interpretation of independence introduced above. In the next subsection, I will discuss some of the challenges that each of these two distinct types of approaches faces.

5.4.1 Measures of independence: A-posteriori and A-priori approaches

The various approaches that have been proposed to define and measure the level of independence across models can be divided into two main families: a posteriori approaches and a priori approaches. Arguably, both approaches can be thought of as attempts to measure inter-model dependencies, that is how dissimilar climate models are from one another in their uncertain assumptions and idealizations about the target system. However, as we will see, each approach has its own set of considerable challenges.

[19] An other objection that has been raised against the error plus truth hypothesis that is worth mentioning is that this hypothesis is incompatible with any concept of "internal variability" (Houghton 2014, 2306; Bishop and Abramowitz 2012, 12; Abramowitz 2019, 96). Indeed as Abramowitz et al. (2019, 96) remark, '[i]f we wish to consider ensemble simulations where unpredictability or aleatory uncertainty is an inherent part of the prediction, we no can longer expect that the system might be entirely deterministically predictable. [...] In these cases we accept that some component of the observational data is inherently unpredictable, even for a perfect model without any epistemic uncertainty'.

Under a posteriori approaches (second interpretation of independence), 'the proximity of GCMs results or of their errors is used to quantify a posteriori their interdependencies' (Boe 2018, 2772). A posteriori approaches can be further divided into ones that assume that the level of dependence depends on the amount of divergence of their outputs independent of observations and ones that assume that it depends on the correlation of model errors (so the latter rely on observations).

Abramowitz and Gupta (2008)'s measure of independence, for instance, belongs to the former. Under their account of independence, the closer the models' outputs are under similar input and initial conditions the more dependent they are considered to be. Several objections have been raised against measures of independence that are based merely on the divergence of outputs independent of observations (such as Abramowitz and Gupta (2008)'s measure). According to Annan and Hargreaves (2017, 213), 'this approach has the potential weakness that models that agree because they are all accurate will be discounted, relative to much worse models, without any allowance being made for their good performance relative to reality.' However, Abramowitz and Gupta (2008, 3-4) do concede that in order 'to choose the best model ensemble, we must consider both the independence *and* performance of potential ensemble members' and that 'choosing model weights for an ensemble is then a process of deciding on a performance measure (or aggregation of performance measures) and then using a weight description that values performance and independence in an appropriate ratio.' [20] So this does seem to be a possible reply to Annan and Hargreaves's concern. Abramowitz et al. (2019)'s objection is stronger, however. They argue that 'inter-model distances alone in the absence of observational data are an incomplete proxy for model independence'. (95) To illustrate their point they consider the following example.

If models' results that are spread around observational estimates should be considered to be independent (even if their outputs are similar), as Abramowitz et al. (2019) argue, and measures of independence that rely on inter-model distances alone in the absence of observational data cannot account for this,

[20] This is of course is easier said than done.

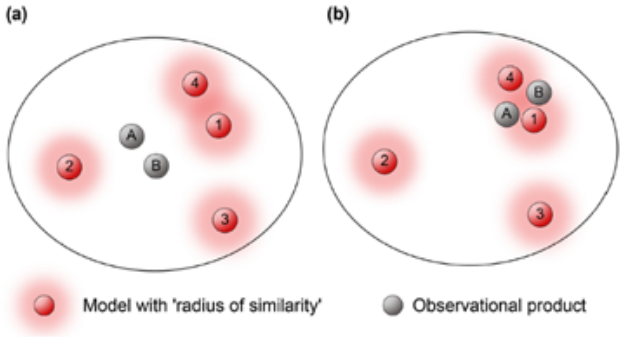

FIGURE 5.3: 'A two-dimensional projection of an inter-model distance space, showing different models and observational estimates, with a radius around models that could be used to determine model dependence. The radius around observations might be related to the uncertainty associated with a given observational estimate. Panels (a) and (b) illustrate how the relative position of observational data sets in this space could complicate this definition of model dependence.' (ibid., 96)

then it seems that one shouldn't use this measure of independence regardless of whether one also takes into account of the models' performance in the weighting process.

Other a posteriori approaches that have been proposed assume that the level of dependence depends on the level of model error covariance or error correlation (Collins et al. 2010, Bishop and Abramowitz 2013). As Abramowitz (2019, 95) notes, these approaches have 'the advantage that "error" only reflects deviations from an observational product (rather than similarity in model outputs per se)' and hence it is perhaps more reasonable to assume that 'differences in the structure of error between models are likely to reflect differences in the sections of model representation that are not tightly constrained by observations'. However, several objections have also been raised against these a posteriori measures of independence. According to Pirtle et al. (2010, 354), *all* a posteriori measures of independence 'essentially treat models as black boxes, ignoring the causal reasons for disagreement between models. It is possible that two models could agree with respect to outputs despite their having different causal assumptions, but such a result, using this approach, would falsely indicate model "dependence," because these models would yield the same output despite the fact that

they make different and possibly conflicting claims about the underlying mechanisms'. Similarly, Annan and Hargreaves (2017, 218) think that '[p]airwise similarity between model outputs may arise through convergence of different approaches to understanding the climate system, and not merely through copying of ideas, and this would not indicate any dependence as defined here. [...] We do not believe that coincidentally similar behaviour should be penalised by downweighting of these models, as it may represent a true "emergent constraint" on system behaviour'. And Abramowitz et al. (2019, 98) worry about the sensitivity of a posteriori dependence measures to the choice of variable, constraining observational data set, metric, time period and the region chosen'.

The a posteriori approaches, discussed above, are seen as pragmatic approaches to quantify inter-model dependencies – the hope with that approach is that the proximity of models' results or model error correlations are good proxy measures for model interdependencies i.e. the similarities in the way the models represent the world and its causal structure.[21] However, all the objections mentioned above raise serious doubts as to whether a posteriori approaches to quantify inter-model dependencies are really fit for purpose. In light of this, some scientists argue that inter-model dependencies should be assessed using a priori approaches instead, where 'the independence of models is judged a priori, based only on the knowledge of their codes, and not of their results.' (Boe 2018, 2772)

A priori approaches are still very much in their infancy, however. A very basic a priori approach is the "institutional democracy" proposed by Leduc et al. (2016). Under this approach models that come from the same institution (i.e. the same modelling center) are assigned less weight.[22] The motivation behind this approach is that '[c]limate models developed within a given research group

[21] But as Abramowitz et al. (2019, 93) point out '[f]or those process representations where models exhibit high fidelity (i.e. where there is sufficient observational constraint to ascertain this), models should be expected to agree in their representation [...] It is only in the cases where there is insufficient observational constraint to diagnose such an epistemic departure, or those where no model can avoid one, that models should provide independent process representations.'

[22] To give a bit more detail, Leduc et al. suggest a weighing technique that gives half weight to a model that agrees with another model from the same centre. So they effectively suggest to'[disregard] one model per pair when an agreement is found' (Leduc et al. 2016, 8310), but not when it is not found. However under a truly a priori approach, agreement across models should play no role in the assessment of inter-model dependencies, so Leduc et al.'s weighing technique seems to be based on a mix of a prioiri and a posteriori considerations.

or institution are prone to share structural similarities' (ibid. 8301) and hence institutional democracy could be used a proxy for measuring inter-model dependences. However, many have found the institutional democracy approach unsatisfactory, since models can share many similarities despite not being from the same modelling centre and hence 'deciding whether or not two GCMs are independent based on their institutions is just a first step. A better knowledge of how code similarity impacts GCMs results is needed to go forward' (Boe 2018, 2772).

Annan and Hargreaves (2017) propose a general account of independence that is determined a priori in terms of the anticipated outputs of the models. According to them two models should be considered independent if a researcher's subjective belief about a possible outcome of one of the models in the ensemble is not affected by learning an output of the other model. However, this assessment of independence is extremely subjective and they only show how it is supposed to work in cases where all the researcher knows is the model's institution.[23]

Boe (2018) has recently proposed to use the number of shared components by GCMs as a proxy for model independence.[24] But he himself acknowledges that this approach 'is still crude and has some limits' (ibid., 2777). For a start, determining whether or not two components are different is not a trivial exercise and is bound to be rather subjective. Indeed Boe relies on the version numbers of the GCMs' components to determine whether two components are different, but as Abramowitz notes 'it is unlikely that the approach to version numbering is consistent across modelling centres, meaning that two components might be very different even if they share a major version number, or vice versa.' (Abramowitz 2019, 94) Another issue that Boe himself points out is that 'the impact of tuning is not considered. Some components may be considered "identical" in this work but use different parameters, which may be a source of important differences. A better documentation of tuning in GCMs would be necessary to go

[23]The downgrading technique they propose to downgrade models that are not deemed to be independent is also based on highly subjective considerations and seems rather arbitrary to me.

[24]Each GCM is characterized by its four key components: atmosphere, ocean, land surface, and sea ice models. Some GCMs may share one or more components, but they may nonetheless use different values for some parameters depending on the tuning strategy that is used or they may also use different versions of that component (where different versions may include substantially different structural assumptions).

further' (Boe, 2777). Yet an other problem is that it is not at all clear how one should select or weigh models based on this measure of independence. As Boe remarks, the simplest approach might be to simply forbid component replication or perhaps to accept a certain level of component replication. However, that would seem to go too far since 'even if GCMs with replicated components are not independent, they are not totally identical.' (ibid., 2777) Boe tentatively suggest a "component democracy" approach, 'whereby each different component would be given the same overall weight in the ensemble. The weight of a GCM would be the combination of the weights corresponding to each of its components' or the 'approach proposed by Annan and Heargreaves (2017) to derive independence weights, but at the replicated component level rather than the group level.' (ibid., 2777) However, these are indeed tentative suggestions. Finally as Abramowitz (2019, 94) remark 'Boé's approach quickly becomes difficult and time consuming for large ensembles such as CMIP, given the lack of transparency regarding precisely what constitutes different models and the role of tuning.' And furthermore 'shared history as it pertains to dependence should only include process representations that are not tightly observationally constrained (so that Navier–Stokes equations might not represent dependent process treatment, for example)' which might further complicate things.

Overall, although a priori approaches to measure inter-model dependencies may intuitively seem more appropriate, they clearly also face considerable challenges. Indeed, there is currently no scientific consensus on how to measure inter-model dependencies.

5.4.2 Why such a strong focus on independence?

In the midst of this search for a measure of independence that can satisfactorily capture how dissimilar climate models are from one another, there is a substantial and yet undefended assumption that is often implicitly and sometimes explicitly made in the literature. It is the assumption that the more dissimilar models are from other models in an ensemble, the greater the confidence we should have in the models' consensus. For instance, when Pirtle argues that 'increased confidence requires an account of independence within each set of

models' and that 'the authors would strengthen their case for increased confidence with a clear account of what these six models share that would make them useful for this study, and an argument for independence despite this overlap', he is implicitly making the assumption that independence across models is able to give one epistemic warrant. So is Boe (2018) when he states that 'At the core of the multi-model approach lies the basic idea that if the results of an additional model B are close to the ones of a model A, then our confidence in the results of A is reinforced. Obviously, it is only true insofar as A and B are not near identical.' And so are Leduc et al. (2016, 8302) when they state that 'Agreements between climate change projections from several models are often interpreted as predictors of confidence [...], but such an inference is difficult to defend without any robust measure of model independence'.

However, this is not at all a trivial assumption. For although it is intuitively clear that if you look at the same thing over and over again you are not going to build any confidence, it would be a fallacy to infer from this that the converse is equally intuitive. That is, is there any reason to believe that if GCMs that are not identical all indicate the truth of a hypothesis this should automatically raise our confidence in that hypothesis as Boe suggests? And in particular is there any reason to believe that the more dissimilar the GCMs are from one another, the more confidence we should have that their consensus is epistemically significant? To explore the answer to the latter question in particular, it will be helpful to go back to some of the arguments for the epistemic import of model consensus that I have discussed in this chapter and see whether they could somehow help us justify why greater dissimilarity across models should raise one's confidence in their consensus. I will start with the two arguments by Parker, discussed in Section 5.2 and I will then briefly turn to Winsberg's argument (although I should really say Schupbach's), discussed in Section 5.3. Finally, I will explore the possible connection between climate scientists' implicit assumption that the more dissimilar models are from other models in an ensemble, the greater the confidence one should have in the models' consensus, and philosophers' (such as Kuorikoski et al. and, arguably, Lloyd too) attempt to justify the epistemic import of model consensus on the basis of some notion of

probabilistic independence.

As discussed in Section 5.2, one of the arguments that Parker (2011) considers as a possible attempt to justify why agreement across models should substantially increase one's confidence in the common result is the following:

1. It is likely that at least one simulation in this collection is indicating correctly regarding hypothesis H.

2. Each of the simulations in this collection indicates the truth of H.

$\therefore$ It is likely that H.

Recall that the problem with this argument is that it seems rather hard to justify that the adequacy condition (premise 1) is met by today's multi-model ensembles, for as Parker argues, neither of the two possible approaches to justify this premise (one that focuses on ensemble construction and one that focus on ensemble performance) is successful. However, one might think that the extent of dissimilarity across models in an ensemble is relevant here. Indeed one might attempt to justify why greater dissimilarity across models in an ensemble should lead to greater confidence in their consensus by relying on the following revised version of Parker's argument above:

1. Greater dissimilarity across models increases the likelihood that at least one simulation is indicating correctly regarding hypothesis H.

2. Each of the simulations in this collection indicates the truth of H.

$\therefore$ Greater dissimilarity across models increases the likelihood that H.

As for Parker's original argument, there seem to be two possible approaches to justify premise 1: One that focuses on model construction and one that focuses on ensemble performance. On the model construction approach, one would argue that the greater the dissimilarity across models, the greater the current scientific uncertainty about how to represent the climate system (for purposes of the predictive task at hand) sampled by an ensemble and hence the more likely that at least one simulation produced in the study is indicating correctly

regarding H. However, why one should accept this argument is far from clear. Without any careful considerations as to the extent to which various dissimilarities are relevant for spanning current scientific uncertainty about how to represent the climate system (for purposes of the predictive task at hand), why should we assume that the more dissimilar an ensemble is, the more it must span current scientific uncertainty? For instance, two models might be quite similar in most respects and yet involve different parameterizations schemes for highly uncertain processes (such as, for instance, cloud formation). Although we might judge these models to be rather similar overall they might arguably span more scientific uncertainty about how to represent the climate system than two models that we might judge to be less similar overall but that do not involve different parameterizations schemes for such highly uncertainty processes. Or perhaps still, two models might be dissimilar in many respects that are nonetheless believed to be irrelevant for the particular purpose at hand (i.e. discerning whether or not H is correct). Although we might judge these models to span more scientific uncertainty than two models that are more similar overall, we might think this greater dissimilarity is irrelevant to the likelihood that at least one simulation is indicating correctly regarding H. That is, considerations on dissimilarity across models do not on their own seem to be sufficient for assessing the extent to which an ensemble samples current scientific uncertainty, and neither do considerations on the extent to which an ensemble samples current scientific uncertainty (independently of the particular purpose at hand) seem sufficient for assessing the likelihood that at least one simulation is indicating correctly regarding hypothesis H. Hence on the model construction approach it seems hard to adequately justify premise 1.

One the performance approach, one would argue that the greater the dissimilarity across models the more justified one is in citing the ensemble's past reliability with respect to H-type hypothesis as evidence that it is likely that at least one of its simulations is indicating correctly regarding this particular H. Recall that one of the worries that Parker had with respect to the performance approach has to do with 'the ad hoc nature of the tuning process' which coupled with ' the fact that today's climate models are far from perfect in their representation

of the climate system' means that 'it cannot be assumed that the performance of a tuned climate model with respect to as-yet-unseen data will be similar to its performance with respect to the data to which it is tuned'. Can independence perhaps alleviate this worry and in so doing provide a justification for premise 1? I do not think so. If the worry from tuning stems from the fact the models are far from perfect in the representation of the climate system (and hence 'because of significant errors elsewhere in the model, parameter values that give the best model performance might be noticeably different from measured values—if a clear physical interpretation of the parameter can be given at all' (Parker 2011, 588)), then it is very unclear why greater dissimilarity across models on its own can address this concern. Indeed it seems to me that the more dissimilar models are in their representation of the climate system, the more reasons to doubt that all of them can be relatively accurate representation of the climate system and the more justified the worry from the ad hoc nature of the tuning process. Hence on the performance approach it also seems hard to adequately justify premise 1.

Let us now look at an other argument that Parker (2011) considers as a possible attempt to justify why agreement across models should substantially increase one's confidence in the common result :

1. e warrants significantly increased confidence in predictive hypothesis H if $p(e|H) >> p(e|\neg H)$.

2. $e =$ all of the models in this ensemble indicate H to be true.

3. The observed agreement among models is substantially more probable if H is true than if H is false; that is, $p(e|H) >> p(e|\neg H)$.

$\therefore$ e warrants significantly increased confidence in H.

As discussed in Section 5.2, Parker argues that premise 3 is rather hard to justify as far today's multi-model ensembles are concerned. One of the reasons that she gives for worrying that models might all indicate the truth of a hypothesis despite the hypothesis being false is that 'when it comes to features and processes that are represented, different models sometimes make use of similar idealizations and simplifications.' The worry here is that models might tend to indicate

the truth of a hypothesis independently of whether the hypothesis is true because they share similar idealizations and simplifications. This is, arguably, also one of the very concerns that is currently driving climate scientists to search for a satisfactory independence metric; one might therefore think that we have finally found where dissimilarity across models might play an epistemic role. The idea here might be something like the following: if we can show that models in an ensemble do not involve many similar idealizations and simplifications, then we might be able to alleviate the worry that models agree merely because they make similar idealizations and simplifications – hence we might be in a better position to justify premise 3.

In this case one might try to justify why greater dissimilarity across models in an ensemble should lead to greater confidence in their consensus by relying on a revised version of Parker's argument above:

1. e warrants significantly increased confidence in predictive hypothesis H if $p(e|H) >> p(e|\neg H)$.

2. $e =$ all of the models in this ensemble indicate H to be true.

3. If the models in this ensemble share few idealizations, simplifications and uncertain factual assumptions then the observed agreement among models is substantially more probable if H is true than if H is false; that is, $p(e|H) >> p(e|\neg H)$.

$\therefore$ If the models in this ensemble share few idealizations, simplifications and uncertain factual assumptions e warrants significantly increased confidence in H.

But a little reflection shows that things are not as straightforward as they seem. As Parker (2006, 363) notes, climate models in a multi-model ensemble 'often incorporate conflicting assumptions about what the climate system is like'. And, arguably, the more dissimilar models are, the more conflicting assumptions one should expect them to incorporate. But if this is right then it is seems to me that having highly dissimilar models in an ensemble merely replaces one worry

with another. For although we can now worry less that models agree merely because they share similar simplifications and idealizations, we now have to worry about why models agree on a result despite making conflicting assumptions about what the climate system is like. In other words, given that the models make conflicting assumptions about the climate system and hence 'the models are [...] incompatible with respect to ontology' (ibid., 364) why should we expect that the models are more likely to agree regarding the truth of a hypothesis on the assumption that the hypothesis is true, rather than on the assumption that the hypothesis is false? If anything, the knowledge that models agree despite making incompatible assumptions about the target system might suggest that the models are agreeing for reasons that are independent of what the climate system is like.

Consider now Winsberg's argument that 'whether or not an ensemble of models is a good candidate for lending strong support for a hypothesis via [robustness analysis] depends almost entirely on the extent to which the set of models suffices for ruling out competing hypotheses.' As argued in Section 5.3, there are many reasons to doubt that Winsberg's attempt to rely on Schupbach's account of RA to justify when agreement across climate models can lend further support to a climate hypothesis works, since despite what Winsberg suggests it doesn't seem that one is ever in a position to formulate an adequate target explanation and rival explanation for the models' common result that satisfy Schupbach's conditions of ERA diversity whenever the models in an ensemble make incompatible assumptions about a target system and the hypothesis we are interested in confirming concerns that target system. But leaving aside the objections I raised in Section 5.3, (in other words supposing that my objections can be somehow be responded to and hence that Winsberg is right in so far consensus across models can lend further support to a hypothesis by allowing us to rule out competing explanations for that result) then clearly greater dissimilarity across models without any detailed account for why those dissimilarities are relevant for ruling out competing hypotheses (if indeed they are relevant) is not going to help one assess the extent to which an ensemble is ERA diverse with respect to a target hypothesis. Hence the idea that the greater dissimilarity

across models, the greater the epistemic import of their consensus can certainly not be defended on the basis of Schupbach's account of ERA diversity.[25]

Finally, it is worth briefly exploring the possible connection between climate scientists' implicit assumption that the more dissimilar models are from other models in an ensemble, the greater the confidence one should have in the models' consensus, and philosophers (such as Kuorikoski et al. and arguably Lloyd too)'s attempt to justify the epistemic import of model consensus on the basis of some notion of probabilistic independence. As discussed in Section 4.2.3 and Section 5.2, by relying on Fitelson (2001)'s account of confirmational independence, one could offer the following argument for why agreement across climate models should increase one's confidence in a hypothesis:

1. If R_1 and R_2 (which are the results of GCM_1 and GCM_2 respectively) individually confirm a hypothesis H and are confirmationally independent regarding H, then $c(H, R_2\&R_1) > c(H, R_1)$, and $c(H, R_2\&R_1) > c(H, R_2)$.

2. R_1 and R_2 individually confirm a hypothesis H.

3. R_1 and R_2 are confirmationally independent regarding H.

$\therefore$ $c(H, R_2\&R_1) > c(H, R_1)$, and $c(H, R_2\&R_1) > c(H, R_2)$.

As discussed in Section 5.2, premise 3 is particularly troubling as far as today's climate model ensembles are concerned. One important reason for this is that current climate models often share similar idealizations, simplifications and also uncertain factual assumptions. As argued extensively in section 4.2.3, when this is the case it is unreasonable to assume their results to be confirmationally independent regarding a hypothesis. One might therefore think that climate scientists' implicit assumption that the more dissimilar models are from other models in an ensemble, the greater the confidence we should have in the models' consensus, is closely connected to this reason for worrying about the validity of

[25]Related to this last point, O'Loughlin (2021), who in contrast to me believes that Winsberg 'convincingly argues that ERA can be applied to climate models' (36) argues that 'because climate scientists may engage in robustness inferences that are not focused solely on pinning down the value of a climate variable and that do not include the elimination of competitor hypotheses, we should be critical of the notion that ERA applies generally across all cases of RA in climate science'. (37)

premise 3 of the above argument. The idea here would have to be something like the following: the more dissimilar models are from other models in an ensemble, the more reasons for believing that premise 3 is justified and hence the more reasons for accepting the above argument's conclusion.

There are, however, at least three problems with this idea. The first is an obvious one. Models' results are either confirmationally independent regarding a hypothesis or they are not so. That is, confirmational independence is not a matter of degrees. Hence, greater dissimilarity across models despite knowing that the models still share some idealizations, simplifications and uncertain assumptions is simply not enough to dismiss our worries about the validity of premise 3. The second, perhaps less obvious, problem with this idea is that the plausibility of premise 3 also depends on what the hypothesis H is all about. Suppose, for instance, that $H = R_T$ is the hypothesis that the models' common result R is instantiated in the target system. In this case premise 3 would require R_1 and R_2 to be confirmationally independent regarding R_T. But as argued in Section 4.2.2, the assumption that the models' results R_1 and R_2 are probabilistically independent conditional on R_T (and $\neg R_T$)[26] is implausible whenever the models under consideration share a set of substantial assumptions about the target system despite differing in all other assumptions. But climate models will inevitably share a large set of substantial assumptions about the target system (e.g. all process representations that are tightly observationally constrained such as, for instance, the Navier-Stokes equations). Hence, it is very unclear why one would have any reason to accept premise 3 in this case, no matter how otherwise dissimilar the models might be from one another. What I take this to illustrate is that the fact that distinct models may share different idealizations, simplifications, and uncertain factual assumptions may in fact be altogether irrelevant for supporting premise 3, depending on the nature of the hypothesis H one is trying to confirm. A final problem with this idea is that there may be cases in which the models under considerations might make distinct incompatible assumptions about the target system and one may have reasons to believe that at most one amongst these

[26]Recall that according to Fitelson (2001) this is a sufficient condition for R_1 and R_2 to be confirmationally independent regarding R_T.

assumptions is adequate, even if one lacks the knowledge to determine which one (e.g. one such case might be when the models involve distinct parametrizations schemes for a given process). As argued in Section 4.2.3, whenever this is the case, it is hard to see why it would be reasonable to assume that conditional on the hypothesis H being correct (whatever H may be), the models' results are probabilistically independent. As discussed in that section, on the assumption that one knows that H is correct, it is unreasonable to suppose that learning that the first model gives result R should not change one's degrees of belief that the second model will give result R. This is because learning that the first model gives result R should give one further reasons to suppose that the assumption in the second model is inadequate, which should reasonably decrease one's degrees of belief that the second model will give result R. Hence, this may be an additional reason for worrying that greater dissimilarity across models per se is irrelevant for dismissing our worries about premise 3 of the above argument.

All in all, it is hard to see why greater dissimilarity across models in an ensemble is relevant for the assessment of the epistemic import of model consensus. This suggests that this frenetic search by climate scientists for a measure of independence that can satisfactorily capture how dissimilar models are from one another not only faces many challenges (some of which I discussed in Section 5.4.1) – it is also misguided.

The assessment of the epistemic import of climate model consensus, and more generally the interpretation of climate models' results, is an extremely challenging problem, one that scientists are trying very hard to deal with; however, as I have suggested in this section, they have not necessarily done so in the most fruitful way. In light of this challenge, some scientists and philosophers (e.g. Stainforth et al. 2007, Betz 2010, Katzav 2014) have argued that the most we should expect from current climate models is for them to be used as tools for articulating 'possibilities'.[27] Betz suggests that, under this view, 'progress would

[27] Whether we should think of these possibilities as 'real possibilities' is itself a source of debate. Suppose we define a real possibility as a state of affairs that has been demonstrated to be consistent with the relevant background knowledge (this is how Betz (2010, 98) seems to implicitly define a real/verified possibility. However, Betz (2016) later refers to this as an *epistemic possibility* and reserves the term 'real possibility (at time t)' to describe all states-of-affairs whose realizations are objectively compatible with the states-of-the-world at time t). In light of this definition, Betz

consist not in convergence of simulation results but in a proliferation of the underlying models, and of the scenarios they generate' while stressing that the current 'prominent role of GCMs is at least debatable. Sophisticated climate models might actually contribute much less to our foreknowledge than evoked by the IPCC'. Although I am somewhat sympathetic to these views, I also believe that a very important, if not main, role of the IPCC is to be responsive to policy makers' and other decision makers' needs for expert judgment, given all the evidence available (including GCMs' results), even though those judgments may indeed have to involve a substantial degree of subjectivity. Hence, I don't think the IPCC should give up trying to establish confidence in claims about how the climate system actually is (contrary to what some proponents of this view seem to suggest). What I do think, however, is that the interpretation of climate model results is indeed a formidable challenge, one the IPCC must deal with somehow or other, and one we philosophers of science, must think very carefully about so as to help, not hinder, the IPCC in their efforts to *practically* deal with it. Although this chapter has been more critical than constructive, I do hope that some of its criticisms can at the very least steer us away from some inauspicious paths and point us towards more promising ones.

(2010, 96) worries that since climate models incorporate assumptions about the climate system that are known to be strictly false, the total states of affairs they represent cannot be considered real possibilities and hence one is not entitled to assume that their predictions represent real possibilities either. Katzav (2014, 236), on the other hand, defines a real possibility (relative to some time t) as follows: (a) its realisation is compatible with the basic way things are in the target domain over the period during which it might be realised and (b) our knowledge at t does not exclude its realisation over that period. In light of this definition, he argues that 'the models only need to provide us with simulations that represent the basic way the climate system is over the periods in question. And representing the basic way the climate system is over a period of time is compatible with being false to a substantial degree. It only requires representing something like the circumstances that obtain in the system and something like the way in which the system evolves. Plausibly, given the substantial knowledge built into GCMs and given the empirical successes of their simulations, their simulations often provide what is required here' (Katzav 2014, 204). However, in light of Katvaz's less demanding definition of 'real possibility' it is unclear, in my view, why one should care about whether something is a real possibility in the first place.

Part III

The weight of evidence and some proposals for a new IPCC uncertainty framework

Chapter 6

On the weight of evidence: what is it, can we measure it, and why should care about it?

6.1 Introduction

It has often been argued there is an important distinction to be made between the *balance of evidence* 'which is a matter of how decisively the data tells for or against the hypothesis' (Joyce 2005) and the *weight of evidence*, 'which is a matter of the gross amount of data available' (Joyce 2005); and that any satisfactory epistemology should recognize this distinction.[1] However, in my view, proponents of this idea have not always been sufficiently clear as to what, according to them, the notion of the "weight of evidence" actually consists in. The one thing that most advocates of this idea do have in common is that they claim that Keynes (1921) was one of the first to point out this distinction and that what they mean by the "weight of evidence" is roughly what Keynes meant by it. Some give a bit more detail; others leave it as that. Yet this is odd for at least two reasons. First, Keynes (1921) never fully clarified what he meant by the notion of the weight of evidence. Indeed, as we will see in the next section, one can distinguish two rather different ways in which Keynes conceptualized the

[1] Joyce (2005) argues that there is also a third aspect of the evidence (i.e. the *specificity* of the evidence) that any satisfactory epistemology should recognize, but I shall not be talking about this third aspect here.

weight of evidence. Second, Keynes had a particular view about the role probabilities should play in our inferences (i.e. he was an advocate of logical probabilities), one that is arguably rather unpopular these days and that is rejected by most recent proponents of this distinction. In light of all this, it is not clear to me whether advocates of this distinction really are referring to an unequivocal concept of the weight of evidence, one that is sufficiently understood to be adequately characterized. In spite of my doubts, I believe the fact that the notion of the weight of evidence has troubled epistemologists for a long time, together with the fact that – despite much effort – a satisfactory measure of it has yet to be found and is unlikely ever to be found, does tell us something. In particular, I will argue that it tells us something about the limitations of an epistemology that envisions the role of probability to be that of quantifying the degree of belief to assign to a hypothesis given the available evidence.

The structure of this chapter is as follows. In Section 6.2, I will introduce what the weight of the evidence consisted in according to Keynes. We will see that Keynes himself thought that the weight of evidence could be understood in at least two different ways, and that it is often impossible to directly measure Keynes's weight of evidence or compare the weight of different evidential sets, however we choose to understand it. In Section 6.3, I will assess the Bayesian's efforts to account for the weight of evidence. I will argue that the Bayesian has not found an adequate measure of the weight of evidence, and that it is unlikely that any will ever be found, for several reasons. In Section 6.3.1, I will argue that the fact the Bayesian worries about the weight of evidence and yet struggles to provide an adequate response to those worries sheds light on the limitations of an epistemology that envisions the role of probability to be that of quantifying the degree of belief to assign to a hypothesis given the available evidence.

Before I begin, I would like to make a cautionary remark. This chapter may, at first sight, appear to be largely disconnected from any issue that I have been concerned with so far and, in particular, from any practical issues that the IPCC authors may face in their evaluation and communication of uncertainty in their findings. Indeed, not only will this chapter be highly abstract; its conclusions will be mostly negative too. Why, then, take my reader on this apparent detour?

What useful and practical lessons can one possibly take away from such an abstract and negative chapter? Yet there are in fact two reasons why, despite prima facie appearances, this chapter is relevant to the assessment and communication of uncertainty by the IPCC.

First, as discussed in Chapter 1, the current IPCC uncertainty framework includes two uncertainty metrics: confidence and likelihood. And, as seen in Chapter 2, although the IPCC gave more than one reason for including two uncertainty scales, those reasons were not articulated with sufficient clarity and precision, nor (when interpreted with sufficient clarity and precision) did they seem like especially good ones. Notwithstanding all this, if indeed there is an important and meaningful distinction to be made between the *balance of evidence* and the *weight of evidence* – a distinction that any satisfactory epistemology should recognize, as many have argued since Keynes first introduced this notion – then this would seem to provide a firm justification for why the IPCC may indeed want to include two uncertainty scales for the communication of uncertainty: one for the balance of the evidence and one for its weight! Given the rather intuitive force of this distinction, the large literature on this matter, and the possible justification that it could provide for why the IPCC should include two uncertainty scales, I believe a careful discussion of this (infamous) notion of the weight evidence is vital if we are to overcome whatever unquestioned intuitions we may have about this distinction and recognize just how very problematic it is.

Second, as we will see in Chapter 7, there are several proposals for a new IPCC uncertainty framework that significantly depart from the current one, and from each other too. Several conclusions of this chapter will be directly relevant to my assessment of at least two of these proposals: Winsberg's (2018) proposal and Bradley et al.'s (2017) proposal. As I will discuss in Section 7.2, the interpretation of confidence under Winsberg's (2018) proposal seems to (if only implicitly) rely on unquestioned assumptions resulting from the literature on resiliency of credence and the weight of evidence. Some of the conclusions

of Section 6.3 of this chapter will help me assess the tenability of those assumptions. Bradley et al.'s (2017) proposal, which I will discuss in Section 7.5, explicitly refers to Keynes's notion of weight of evidence to motivate the role of confidence under their proposal. However, I will argue that not only is it highly unclear why confidence under this proposal should have anything to do with Keynes's notion of the weight of evidence, which I will discuss extensively in Section 6.2 in this chapter, but that the only possible way in which confidence under their proposal can be understood as expressing something remotely close to the notion of Keynes's weight of evidence renders their proposal conceptually flawed.

In a nutshell: if you do find yourself questioning the relevance of this chapter, please stay with it all the same, as its pertinence will become clear in due course.

6.2 Keynes on the weight of arguments: two unmeasurable concepts

The aim of this section is to give an overview of Keynes's notion of 'the weight of an argument'. But before I do that, it will be helpful to give a quick introduction to Keynes's interpretation of probability. Keynes is an advocate of *logical* probabilities. That is, according to him, probability is a logical relationship between some premises E and a conclusion H: the truth of the premises *entails* some degree of rational degree of belief in the conclusion.[2]

There are two important features of probability resulting from this view that are worth mentioning. First, according to this view, unconditional probabilities are meaningless, since probability is always a relation between some premises and a proposition. In other words, no proposition by and of itself is probable or improbable. Second, probabilities are always objective, in the sense that there

[2]It is worth mentioning, however, that according to Keynes numerical probabilities were (very) special cases of probability, which neither had to be quantifiable nor comparable. Indeed, according to Keynes, only under very specific conditions could probabilities be numerical, such as, for instance, under the conditions of Keynes' own version of the Principle of Indifference (see Keynes (1921, Chapter 4).

is a unique rational degree of belief in a proposition given some premises. In Keynes's words:

> A proposition is not probable because we think it so. When once the facts are given which determine our knowledge, what is probable or improbable in these circumstances has been fixed objectively, and is independent of our opinion. (Keynes 1921, 4)

From this view it follows that whenever distinct individuals disagree about probabilities when faced with the same conclusion and premises, they cannot all be correct (i.e. some of them must have made a logical fallacy). However, since the probability of a proposition is always relative to some premises, and the premises that are selected will always depend on the evidence available to the particular individual at a particular time, distinct individuals *can* rationally assign different probabilities to the same proposition. Whenever distinct rational individuals disagree about the probability of a proposition because they have selected different premises in light of the available evidence, there is no sense in which one probability is more correct than another according to Keynes. All probabilities $Pr(H|E_1), Pr(H|E_2) \ldots$ are correct. For instance, consider a case in which the conclusion of two arguments is the same, and the relevant evidence in one of the two arguments includes and exceeds the evidence in the other. Although the two arguments might have different probabilities, there is no sense in which the probability of the argument which includes additional evidence is more correct than the one which includes less evidence:

> If $Pr(H|E_1\&E_2) = 2/3$ and $Pr(H|E_1) = 3/4$, it has sometimes been supposed that it is more probable that $Pr(H|E_1\&E_2)$ really is $2/3$ than that $Pr(H|E_1)$ really is $3/4$. According to this view, an increase in the amount of evidence strengthens the probability of the probability, or, as De Morgan would say, the presumption of the probability. A little reflection will show that such a theory is untenable. For the probability of H on hypothesis E_1 is independent of whether as a matter of fact H is or is not true, and if we find out subsequently that H is true, this does not make it false to say that on hypothesis E_1 the

> probability of H is $3/4$. Similarly the fact that $Pr(H|E_1\&E_2)$ is $2/3$ does not impugn the conclusion that $Pr(H|E_1)$ is $3/4$, and unless we have made a mistake in our judgment or our calculation on the evidence, the two probabilities are $2/3$ and $3/4$ respectively. (Keynes 1921, 80)

With this very brief overview of Keynes's interpretation of probability, it is time to turn to his notion of the weight of an argument. Before we do that, though, a clarification is in order. As mentioned in the introduction, Keynes's interpretation of probability is not widely shared these days. Hence, one might reasonably worry: if Keynes's notion of the weight of an argument strictly relies on this particular interpretation of probability, then why should anyone who does not share this interpretation care about this notion? In other words, what can one hope to learn from a discussion about what Keynes's notion of the weight of an argument is all about if one does not agree with his interpretation of probability to begin with? In light of this potential worry, I'd like to stress that nothing substantial about my overview of Keynes's notion of the weight of an argument strictly relies on Keynes's own particular understanding of probability. In other words, Keynes's interpretation could be replaced by an interpretation that is more plausible by modern lights without affecting the discussion to come. Having said this, certain features of Keynes's notion of weight of an argument will indeed be affected by the kind of interpretation of probability one is operating with and I will make sure to point those out whenever this is the case.

Below is an often quoted passage in Keynes's *Treatise on Probability* on this notion:

> As the relevant evidence at our disposal increases, the magnitude of the probability of the argument may either decrease or increase, according as the new knowledge strengthens the unfavourable or the favourable evidence; but something seems to have increased in either case, - we have a more substantial basis upon which to rest our conclusion. I express this by saying that an accession of new evidence increases the weight of the argument. New evidence will

> sometimes decrease the probability of an argument, but it will always increase its 'weight'. (Keynes 1921, 77)

As hinted in the above quote, according to Keynes an argument from premises E to conclusion H has, in addition to a probability $Pr(H|E)$, also a weight $V(H|E)$. However, despite this passage is often quoted to allude to what Keynes has in mind with his notion of the weight of an argument, what this concept actually consists in according to Keynes is far from clear; for as Runde (1990) remarks, one can distinguish two conceptually different notions of the weight of an argument in Keynes's *Treatise*.

According to the first conception of the weight of an argument, which following Runde (1990), I call weight$_1$, 'one argument has more weight than another if it is based on a greater amount of relevant evidence' (Keynes 1921, 84). From this conception of the weight of an argument, it follows that the probability of an argument is completely independent from its weight. For if all that matters for comparing the weight of two arguments is the amount of relevant evidence that appears in their premises, then it is clear that one can in principle compare the weight of two arguments without having to know their probabilities. Under weight$_1$, it also follows that the addition of relevant evidence E_1 to the original premises of an argument $H|E$ will *always* increase its weight (i.e. $V(H|E\&E1) > V(H|E)$) independently of whether the probability of the new argument is higher, lower or the same as that of the original argument (i.e. independently of whether $Pr(H|E\&E1) > Pr(H|E)$ or $Pr(H|E\&E1) < Pr(H|E)$ or $Pr(H|E\&E1) = Pr(H|E)$).

For weight$_1$ to be a meaningful concept we must define what it means for some evidence to be relevant. One may prima facie be tempted to say that: E_1 is relevant to H on evidence E if an only if $Pr(H|E\&E_1) \neq Pr(H|E)$ i.e. if and only if the addition of E_1, to data E makes a difference to the probability of H. However, as Keynes recognizes, this definition of relevance is too strict. To see why this is consider a case in which the addition of premise E_1 affects the probability of an argument in one direction exactly as much as the further addition of E_2 affects it in the other. In this case the addition of E_1 would increase the weight of

the argument and the further addition of E_2 would further increase the weight of the argument. But then if we don't think it reasonable for the weight of an argument to be affected differently depending on whether E_1 and E_2 are added successively or conjunctively, the addition of the conjunction $E_1\&E_2$ should also increase the weight of the argument even though it is not relevant in the sense above. In light of this possibility, Keynes attempts to provide a less strict definition of relevance. According to his definition a proposition E_1 is relevant to H on evidence E if and only if there is a proposition E_2 inferable from $E_1\&E$ but not from E such that $Pr(H|E\&E_2) \neq Pr(H|E)$. However, as Cohen (1986) remarks, Keynes's less strict definition of relevance can't do the job he wants it to do, since *any* proposition E_1 whatsoever entails the disjunction $E_1 \vee H$, and $Pr(H|E\&(E_1 \vee H)) \neq Pr(H|E)$.[3] Hence, from Keynes's less strict definition of relevance, it follows that any proposition whatsoever is relevant to an argument and hence will increase its weight. Since this would evidently trivialize the concept of the weight of an argument, Cohen (1986) suggests to tighten the conditions under which some evidence E_1 increases the weight of an argument as follows (in the passage below he is using the standard definition of relevance, i.e. E_1 is relevant to H on evidence E if an only if $Pr(H|E\&E_1) \neq Pr(H|E)$):

> In order to avoid such trivialisation we need to tighten the conditions under which $V(H|E\&E_1) > V(H|E)$. We need to say that this inequality holds if and only if E_1 entails a proposition E_2 that is relevant to $Pr(H|E)$, where no proposition E_3 occurs in E_2 (or in any equivalent of E_2) such that E_1 entails E_3 and, without affecting the relevance of E_2 to $Pr(H|E)$, E_3 can be replaced in E_2 (or in some equivalent of E_2) by a proposition that has no relevance to $Pr(H|E)$. And we can also say that under just these same conditions E_1 will give at least as much weight to $Pr(H|E)$ as E_2 does. (Cohen 1986, 268)

Under Cohen's suggestion it is no longer the case that any proposition whatsoever will increase the weight of an argument. Indeed, notice that we can replace

[3] Since $Pr(E_1 \vee H|H\&E) = 0$, by applying Bayes's theorem it follows that $Pr(H|E\&(E_1 \vee H) = \frac{Pr(H|E)}{Pr(E_1 \vee H|E)} \neq Pr(H|E)$.

E_1 in $E_2 = E_1 \vee H$ with a proposition that has no relevance to $Pr(H|E)$ without affecting the relevance of E_2 to $Pr(H|E)$. Hence, under Cohen's conditions, the fact that E_1 entails the disjunction $E_2 = E_1 \vee H$ is no longer a reason to assume that E_1 will increase the weight of an argument. Hence, I am happy to assume that Cohen's definition of relevance[4] is adequate.

Equipped with this definition of relevance, $weight_1$ seems to be a meaningful concept.[5] However, the following question arises. Can we always compare the weight of two arguments? Keynes himself doesn't think so. According to him in a very large number of cases it is in fact impossible to compare the weight of two arguments:

> Where the conclusions of two arguments are different, or where the evidence for the one does not overlap the evidence for the other, it will often be impossible to compare their weights, just as it may be impossible to compare their probabilities. (Keynes 1921, 78)

Indeed, let me illustrate some of the difficulties encountered when trying to compare the weight of two arguments through the following example. Consider these two arguments: one is from $E_1\&E_2$ to H and the other is from $E_1\&E_3$ to H. In this case, the premises of the first argument do not entail all the premises of the second argument and vice versa. Suppose that $Pr(H|E_1\&E_2)$ differs more from $Pr(H|E_1)$ than does $Pr(H|E_1\&E_3)$. One might think that since E_2 is of greater relevance than E_3, the weight of $Pr(H|E_1\&E_2)$ should be greater than that of $Pr(H|E_1\&E_3$. However, as Cohen convincingly argues, allowing the the extent of a new premise's relevance to enter into comparisons of incremental weight would lead one into a paradoxical situation, one in which the order in which different premises are stated could affect the weight of an argument:

[4]Cohen's conditions under which $V(H|E\&E_1) > V(H|E)$ can be reinterpreted as conditions under which E_1 counts as relevant to the weight of an argument.

[5]It is worth mentioning, however, that an important feature of this concept will be affected by the kind of interpretation of probability one is operating with. Indeed, notice that under Keynes interpretation of probability, $weight_1$ is an objective notion in so far as distinct rational individuals when faced with the same conclusion and premises can't disagree about their views on the weight$_1$ of such argument. This is because under Kaynes' interpretation, they can't disagree about what counts as relevant evidence. Whereas, for instance, under a subjective Bayesian interpretation of probabilities, distinct rational individuals when faced with the same conclusion and premises could disagree about their views on the weight$_1$ of such argument because they *can* rationally disagree about what counts as relevant evidence.

> Suppose a set of evidential items $E_1, E_2, \ldots, E_{100}$ in regard to a hypothesised conclusion H. Suppose too that quite a lot of these items, on their own, ground low probabilities in favour of H, quite a lot ground high probabilities in favour of H, and quite a lot ground intermediate probabilities at varying levels. One way of ordering these items would be to begin with those highly in favour of H, then proceed with those slightly less in favour and so on down, ending up with those highly in favour of $\neg H$. In such a carefully graduated order the extent of the relevance of each new piece of evidence, after the first, would tend to be small. So if the weight of the argument were to be affected by the extent of the relevance of each incremental piece of evidence, as well as by the number of those pieces, the additional effect on the overall weight would be minimal. But, if instead the evidential premisses were ordered so as to alternate as violently as possible between favourable and unfavourable items, the overall effect on the weight would be very different, if extent of relevance was allowed to affect the issue at each incremental step. (Cohen 1986, 271)

In light of this, it seems unreasonable to assume that since E_2 is of greater relevance than E_3, the weight of $Pr(H|E_1\&E_2)$ is greater than that of $Pr(H|E_1\&E_3)$. So using Cohen's own example, although learning that a person has a dangerous hobby (E_2) might be more relevant to the probability that a person will survive to age 65 than learning that the person is a male (E_3), given that the person is a lorry-driver (E_1), one would not be justified in concluding from this that the probability that a lorry-driver with a dangerous hobby will survive to age 65 has greater weight than the probability that a male lorry-driver will survive to age 65.[6]

But then one may wonder: does it follow that these two probabilities should

[6]When I speak of 'the weight of the probability of a hypothesis H given evidence E', what I should really say is 'the weight of the argument from E to H'. However, this flexibility is harmless as long as one remembers that according to Keynes the weight is independent of the probability value (Keynes (1921) himself is pretty flexible in his notation too).

be given the same weight? According to Cohen the answer is again no: the inequality in this case is rejected because no comparisons of this kind are possible, rather than because the true comparison is one of equality. Why is that?

According to Cohen the assumption that the two probabilities have the same weight in this case would have to rely on the acceptance of what he calls 'the principle of equipollence', 'that the members of different families of predicates enhance the weight of an argument equally when they enter relevantly into its premisses' (Cohen 1986, 273). However, Cohen argues this principle is implausible. To convince us of this, he asks us to consider the two predicates 'has a dangerous hobby' and 'has a dangerous hobby and a weak heart'. Under the principle of equipollance it would seem to follow that the weight of an argument must be the same regardless of which of these two predicates enters into the premises of the argument. But this is, arguably, implausible, since having a weak heart is certainly relevant to whether a person survives to age 65 even on the condition that the person is a lorry-driver and has a dangerous hobby. Hence the only way to salvage the principle of equipollence would be to restrict its application to primitive predicates in some appropriately tailored language-system. However, as Cohen remarks, this move 'would introduce a substantial element of linguistic convention into the assessment of weight. The weight of an argument would depend not just on facts about probabilistic relevance but also on which predicates were chosen as primitive and therefore as having no non-trivial entailments' (Cohen 1986, 274). Hence, Cohen concludes that 'unless there is a reason in a particular area of inquiry to suppose that the primitiveness or non-primitiveness of a predicate is unambiguously determined by the facts rather than convention, it looks as though the principle of equipollance cannot be rescued' (ibid., 274). Notice further that if Cohen is right, that is if the principle of equipollance really cannot be rescued, this also means that 'there is no natural unit of weight and the prospects of any non-arbitrary system for measuring weight are very poor' (ibid., 274). Hence, without the help of the principle of equipollance, there is no sense in which we can measure the weight of an argument. All we can do is compare the weight_1 of distinct arguments, and as we have seen above, even this comparison will not be possible in most cases.

Leaving aside considerations of measurability and comparability, there is a much a more pressing question. Why should we *care* about $weight_1$ from an epistemological perspective? Is it really the case that when faced with an argument with greater weight_1 than another, 'we [always] have a more substantial basis upon which to rest our conclusion', as Keynes remarks in the passage quoted at the beginning of this section? As some have argued (Runde, 1990; Feduci, 2010) this doesn't seem to be the case, since the acquisition of more relevant evidence does not necessarily lead to a more a substantial basis on which to rest our conclusion. To see why this is, it will be helpful to turn to Keynes's *second* conception of the weight of an argument.

There is another conception of the weight of an argument to which Keynes alludes in his *Treatise*, and which consists in a substantial departure from weight_1. According to this second conception of the weight of an argument, which following Runde (1990) I will call weight_2, the comparison of the weight of two arguments 'turns upon a balance [...] between the absolute amounts of relevant knowledge and of relevant ignorance respectively' (Keynes 1921, 77) and thus depends on 'the degree of completeness of the information on which a probability is based'[7] (ibid., 345). Weight_2 is a different concept from that of weight_1 and yet Keynes neither says much about it, nor does he acknowledge the distinction in the first place. However, the distinction is certainly there and hence treating weight_1 and weight_2 as one and the same concept, as Keynes does, is wrong.

Under weight_2, comparing the amount of relevant evidence that appears in the premises of two arguments is no longer sufficient for comparing their weights. Since on this account, the weight of an argument no longer depends on merely the amount of evidence that appears in its premises as under the previous account. Rather, it now depends on the amount of relevant evidence *and* the amount of relevant ignorance. Intuitively, the relationship between weight_1 and weight_2 can be understood by the following metaphor by Feduzi:

[7] Runde (1990, 281) actually distinguishes three conceptions of the weight of evidence in Keynes's *Treatise*. Let K_r represent the relevant knowledge and I_r relevant ignorance. According to Runde, Under weight_1, $V(H|E) = K_r$; under weight_2, $V(H|E) = \frac{K_r}{K_r+I_r}$ and under weight_3, $V(H|E) = \frac{K_r}{I_r}$. However, weight_2 and weight_3 are conceptually very similar, in particular if one increases so does the other. Hence, as Runde himself concedes the distinction is conceptually irrelevant.

> If I tell you that 'I have covered twenty miles', you cannot say if I have come very far in my journey. But if I tell you my final destination, you can tell if I am at the beginning or at the end of my journey, or if I am almost there. In the same way, the absolute amount of evidence E one has already acquired (first definition of weight) does not reveal how far one has come in the learning process [...]; but if one knows how much information is relevant to the proposition H, one can say whether the evidence acquired so far is relatively 'scanty', 'complete' or simply 'sufficient' to make a decision [...]. This is because no evidence is itself 'scanty' or 'complete' in the same way as no place can be intrinsically distant. (Feduzi 2010, 343)

Although this metaphor give us an intuitive way to understand the difference between weight_2 and weight_1, it is not at all clear whether weight_2 is in fact a meaningful concept in the first place. Whether or not it is meaningful strictly depends on whether a definition of relevant ignorance is possible. Keynes, himself, does not provide a definition of relevant ignorance so it is hard to know what he really has in mind as far as this concept is concerned. However, Runde (1990, 282) suggests that relevant ignorance should be understood as all those factors of which an agent is 'to a large extent ignorant, but which are relevant to [her] probability estimates.' According to Runde, one is often able to identify such factors. He provides the following example:

> Consider, for example, the proposition r that it will rain two days hence. On the basis of the evidence, namely, certain propositions we take to be true, we may be able to arrive at the probability of r. These evidential propositions take the form of "direct knowledge" in Keynes's account. In practice, we may use certain historical data in the belief that it is the best available, or rely on recent weather forecasts and meteorological reports. We are nevertheless aware of the possibility, in these situations, that better data may be available, or that it may have been an apprentice weatherperson who has been making the reports in recent weeks. And by the same token, we are

> aware that there are usually relevant factors that we have omitted altogether. It is in these senses, I maintain, that we may speak of "relevant ignorance." (Runde 1990, 282)

Some remarks about Runde's 'definition' of relevant ignorance are in order. The first thing to note is that relevant ignorance according to Runde's definition is to a large extent a subjective notion, since the assessment of the extent of an agent's relevant ignorance strictly relies on the agent's ability to determine those factors of which she is ignorant but that are relevant to her probability estimates, and this will clearly vary from agent to agent. The second thing to note is that this definition of relevant ignorance exclusively concerns relevant factors of which the agent is aware. But this raises a question: why should an agent restrict her attention to relevant factors of which she is aware for her assessment of her relevant ignorance? For instance, consider a case in which an agent is not able to identify any particular factor that is relevant to her probability estimates and yet she nonetheless believes that there may be factors which she doesn't know but that are nonetheless relevant to her probability estimates. According to Runde's definition it seems that the agent in this case would have to conclude that she has no relevant ignorance. But this seems a little odd. The agent's belief in the existence of factors of which she is ignorant, but that are relevant to her probability estimates, should arguably be taken into account in her assessment of her relevant ignorance regardless of whether or not the agent is able to identify those factors. Relatedly, Feduzi (2010, 344) distinguishes four different epistemic situation in which an agent might find themselves:

1. The agent knows all the available evidence relevant to some conclusion and knows that she knows all of it.

2. The agent does not know some of the evidence relevant to some conclusion and knows that this is the case.

3. The agent does not know some part of the evidence relevant to some conclusion, does not know that she does not know this part of the evidence,

but knows that there may be some part of the evidence that she does not know.

4. The agent does not know a part of the evidence relevant to some conclusion, does not know that she does not know this part of the evidence, and does not know that she might not know some relevant evidence.

The first situation is one in which there are no relevant factors of which the agent is ignorant. The second is one in which the agent is ignorant of some relevant factors and she is able to identify such factors. The third is one in which the agent believes that there may be some relevant factors of which she is ignorant, but she is not able to identify what they are. And the fourth one is one in which there are relevant factors of which the agent is ignorant, but the agent does not believe this is the case.

The example I gave above is one in which the agent is in the third situation on this list, and according to Feduzi an agent will often find herself in this situation:

> In many cases the decision maker cannot recognize the main features of her ignorance. The decision maker is frequently unable even to imagine factors that could affect the probability of an event. However, I claim that she is always aware of the possibility that there might be relevant factors that she could have omitted altogether. Situation (3) thus represents a choice situation in which the decision maker is not able to recognize relevant factors of which she is ignorant, but she is 'aware of the possibility of being surprised'; she does not 'have in mind' how she is going to be surprised, but she knows that this eventuality is likely to happen. (Feduzi 2010, 345)

Notice that under Runde's definition of relevant ignorance an agent's assessment of her relevant ignorance would be the same under situation 1,3 and 4: in all these situations she would conclude that she has no relevant ignorance. The only situation in which she would conclude that she has some relevant ignorance is in situation 2. Hence if we don't think that is right, that is, if we think that the assessment of an agent's relevant ignorance should be affected by

whether she finds herself in situation 3 on the one hand and situation 1 and 2 on the other, then Runde's definition *must* be revised.[8]

But even if we were to settle on an appropriate definition of an agent's relevant ignorance (regardless of what that might be), to what extent can we compare the weight$_2$ of distinct arguments? Recall that if the principle of equipollance is indefensible, as Cohen argues, the weight$_2$ of an argument is not measurable since neither the amount of relevant knowledge nor the amount of relevant ignorance can be measured in any non-arbitrary way. So, as for weight$_1$, the best we can hope for is to *compare* the weight$_2$ of distinct arguments. However, the comparison of the weight$_2$ of distinct arguments seems to be even more tricky than it was for weight$_1$.

To see why this is, consider a case in which the acquisition of new relevant evidence also has the consequence of increasing an agent's perception of the amount of her relevant ignorance. According to Runde this is a case where the weight$_2$ of an argument might actually decrease despite the fact that one has gained more relevant evidence (contrary to weight$_1$ which always increases with the addition of relevant evidence):

> In terms of weight$_1$ new evidence "will sometimes decrease the probability of an argument, but it will always increase its weight" [...]. The surprising feature of weight$_2$ is that the same conclusion need not follow. New evidence, in other words, may lead to a decrease in weight. To see this, it will be helpful to refer again to [$V(H|E) = \frac{K_r}{K_r+I_r}$][9]: If I_r does not increase by more than K_r, it is clear that weight$_2$ will increase with every increase in K_r. But it is surely possible, in principle, that we may sometimes learn something that leads us to drastically reassess I_r, to revise it upward by more than any increase in K_r. In this case, the accretion of evidence will lead to a decrease in weight. (Runde 1990, 282)

[8]Notice that even though situation 4 and situation 1 are different epistemic situations, if the assessment of relevant ignorance is relative to the agent's *subjective beliefs* about what she does not know, then regardless of how we cash this idea out in its details, an agent's assessment of her relevant ignorance cannot be affected by whether she finds herself in situation 4 or situation 1.

[9]See footnote 5.

However, things are not at all as easy as Runde would like us to believe. This is a case in which *both* the amount of relevant knowledge and the amount of relevant ignorance of an agent have increased, and because of this very fact it is unclear how an agent can determine whether the weight_2 of the argument has increased, decreased or stayed the same. To determine whether it has increased, decreased or stayed the same, the agent would have to determine *how much* the amount of relevant knowledge has increased compared to the original amount of relevant knowledge, and *how much* the amount of relative ignorance has increased relative to the original amount of relevant ignorance. But again if the principle of equipollance is indefensible, as Cohen argues, this doesn't seem to be possible. Hence, clearly, comparisons of the weight_2 of distinct arguments are even more difficult to come by than comparisons of the weight_1 of distinct arguments. In particular, in contrast to weight_1, whenever the acquisition of new relevant evidence also has the consequence of increasing an agent's perspective of the amount of her relevant ignorance it seems impossible to determine how this should affect the weight_2 of an argument.

In light of the discussion above, it is clear that weight_2 is a considerably more difficult concept to grasp than that of weight_1 since its nature will depend on how we choose to define an agent's relevant ignorance, and that choice is clearly not as straightforward as Runde seems to suggest. Furthermore, in contrast to weight_1, the assessment of the weight_2 of an argument depends on an agent's own awareness of her relevant ignorance, which can vary considerably from agent to agent. Finally, we have also seen that comparisons of the weight_2 of distinct arguments seem to be even more challenging to come by than comparisons between the weight_1 of distinct arguments.

However, despite the fact that weight_2 is a considerably harder notion to grasp than weight_1 if as Feduzi remarks 'the absolute amount of evidence E one has already acquired (first definition of weight) does not reveal how far one has come in the learning process', then it seems that if we are to be at all persuaded by Keynes's claim that a greater weight implies that 'we have a more substantial basis upon which to rest our conclusion', weight_2 is evidently a more appropriate conception of the weight of an argument than weight_1.

Before concluding this section, it is worth mentioning that although Keynes did think that our decisions in light of the available evidence E relevant to a hypothesis H should be affected by both the probability of an argument $Pr(H|E)$ and its weight $V(H|E)$, he himself struggled to understand how all this was supposed to work:

> In the present connection the question comes to this - if two probabilities are equal in degree, ought we, in choosing our course of action, to prefer that one which is based on a greater body of knowledge? ... The question appears to me to be highly perplexing, and it is difficult to say much that is useful about it. But the degree of completeness of the information on which a probability is based does seem to be relevant, as well as the actual magnitude of the probability, in making practical decisions. Bernoulli's maxim, that in reckoning a probability we must take into account all the information which we have, even when reinforced by Locke's maxim that we must get all the information we can, does not completely seem to meet the case. If, for one alternative, the available information is necessarily small, that does not seem to be a consideration which ought to be left out of the account altogether. (Keynes 1921 345-46)

This ends my discussion of Keynes's notion of the weight of an argument. In the next section I will turn to what the Bayesian has had to say about this notion, since Keynes first introduced it.

6.3 The Bayesian on the weight of evidence

In the previous section, we have seen that what the notion of the weight of an argument (which from now on I will refer to as 'the weigh of evidence', in line with how people refer to it nowadays) actually consists in, according to Keynes, is not clear, since he offers at least two distinct conceptions of it. We have also seen that any attempt to directly measure the weight of evidence (regardless of whether we understand it as weight_1 or weight_2) does not seem to be possible. We can at best compare the weight of different evidential sets and even this

doesn't seem to be possible in most cases. But regardless of the elusiveness surrounding this concept, many Bayesians seem to agree with Keynes: the weight of evidence is not a meaningless concept and however we choose to understand it, it is not represented in an agent's credences. In light of this recognition, the Bayesian has gone to considerable length to try to show that they can account for the weight of evidence in other ways. In this section, I will assess the extent to which these Bayesian's efforts are successful. First, though, I need to make an important clarification.

Several philosophers (e.g. Popper 1959, 424; Good 1985, 267; O' Donnell 1989, 76; Roussous 2020, 197; and many more) have claimed that the first person to introduce the notion of the weight of evidence was not Keynes (1921) but Peirce (1878) many years earlier. But as Kasser explains,

> The situation actually involves an embarrassment and a confusion of riches, however. Peirce formulates two quite distinct notions of weight of evidence, each of which has been influential. One anticipates Keynes's conception of the weight of argument, first broached in his 1921 A Treatise on Probability. Peirce develops this sense of weight as part of a critique of conceptualist (or, as we would now say, Bayesian) approaches to probability. But Peirce also develops a conception of weight of evidence that favors conceptualism, and this has been picked up by Bayesians like Good. Peirce goes to considerable trouble to distinguish the two notions, but almost all commentators have either conflated the two notions or ignored one in favor of the other. (Kasser 2016, 629)

Indeed, two very different understandings of the weight of evidence can be traced back to an 1878 article by Peirce "The Probability of Induction". In that article Pierce introduces two distinct notions. One of these is supposed to measure the gross amount of evidence and seems to anticipate Keynes's conception(s) of the weight of evidence (Peirce himself did not give it a name, however). The other notion Peirce introduced in that article and to which he referred as the

weight of evidence, is supposed to measure something like the *balance* of evidence in favour of a hypothesis compared to another. This notion of the weight of evidence has become associated with I. J. Good who was one of the first to formalize it and is considered to this day to be a useful notion in Bayesian analysis (see, for instance, Fairfield and Charmann (2017)). According to Good (1950, 1985) the weight of evidence in favour of a hypothesis H_i compared to a rival hypothesis H_j is proportional to the logarithm of the likelihood ratio:[10]

$$WOE(H_i : H_j) = \log_{10} \frac{P(E|H_i)}{P(E|H_j)}.$$

I mention this because these two notions of the weight evidence (Keynes's and Good's) are in fact two completely distinct notions that should neither be conflated nor, in my view, seen in competition with one another. Indeed, as Joyce (2005, 165) remarks Good's measure above is really a measure 'of evidential relevance that compare[s] balances of total evidence irrespective of weight. Since the values of these measures can remain fixed even as the volume of data increases, they do not capture the weight of evidence in the sense Keynes had in mind'. Despite this, as Kasser notes 'an attempt to characterize one kind of evidential weight has often been criticized as a misguided attempt to measure the other' (Kasser 2016, 641). Hence, I just want to make it clear that in what follows I am exclusively focusing on Keynes's notion of the weight of evidence and whether or not the Bayesian can account for this notion.

An example that is often used to motivate the idea that the weight of evidence is not reflected in an agent's credences is what Popper called 'the paradox of ideal evidence' (Popper 1959, 425-7) even though it is not a paradox in any sense of the term. It goes as something like the following:

> John is presented with a coin and he is asked to assign a probability to the proposition H that it will come up heads next time it is tossed.

[10]The logarithmic scale is in many cases considered to be more natural than a linear scale for measuring sensory inputs (e.g. sound which is measured in decibels). For similar reasons, logarithm scales are also thought to facilitate the assessment of perceptions of uncertainty in probabilistic inference (see e.g. Peirce (294) Good (1985, 255)), which partly explains why Good's notion of the weight of evidence is to this day considered to have practical value in Bayesian inferences (see, for instance, Fairfield and Charmann (2017).

> He doesn't know whether the coin is fair or whether it is biased towards heads or tails. So in light of his ignorance, he assigns a probability of $1/2$ to H i.e. his subjective prior in H is $Pr(H) = 1/2$. He is then allowed to toss the coin a thousand times and he gets about 50% heads and 50% tails. Call this evidence E. The probability that he assigns to the proposition H that the coin will land heads next time it is tossed is still $1/2$ i.e. $Pr(H|E) = Pr(H) = 1/2$.

Popper is troubled by this example: after observing the coin being tossed a thousand times, John has a lot more evidence concerning the proposition H that the coin will come up heads next time it is tossed, and yet this is not reflected in John's credence in H. i.e. $Pr(H|E) = Pr(H)$. In other words, the weight of evidence is not reflected in John's credence for the outcome of the coin next time it is tossed.

The standard reply from the Bayesian to Popper's concerns is the following: true, the weight of evidence is not reflected in John's credence in H, but it is nonetheless reflected in the *resiliency* of his credence in H, so nothing to worry about! This reply goes all the way back to Jeffrey (1965, 196) who notes that although John prior to seeing the evidence (call him John$_{\neg E}$) and John after seeing the evidence (call him John$_E$) assign the same probability to the proposition H that the coin will come up heads the next time it is tossed, they nonetheless 'assign different values to any proposition $A(n)$ that asserts, concerning $n \geq 2$ distinct tosses, that all of them yield heads. To any such proposition [John$_E$] assigns the value $1/2^n$; but to the same proposition [John$_{\neg E}$] must assign a higher value, if you hope to learn from experience' (Jeffrey 1965, 196). What Jeffrey is essentially pointing out here is that John$_E$'s credence in H is not going to change very much in the face of new evidence, regardless of what that evidence is (e.g a long series of tosses all yielding heads) and that is why he assigns the (approximate) value $1/2^n$ to any proposition $A(n)$. On the other hand, John$_{\neg E}$'s credence for H will increase each time he sees a coin toss yielding heads, and hence why he will assign a probability greater than $1/2^n$ to any proposition $A(n)$. In other words, as Skyrms (1977, 707) puts it 'the ideal evidence has changed not the

probability of [heads] on toss a, but rather the resiliency of the probability of [heads] on toss a'.

The reason why John$_E$'s credence in H is more resilient than John$_{\neg E}$'s is not obscure. Although evidence E has not affected the probability that John assigns to the proposition H, it has affected John's credences in other hypotheses. This is how Skyrms puts it:

> [I]n the ignorance situation the second order probabilities are spread out all over the spectrum for $Pr(heads) = x$ [though we may plausibly assume that the second-order probability weighed average for values of $Pr(heads) = 1/2$; i.e., the second-order expectation, is $1/2$]. In the "ideal-evidence" situation, the second-order probabilities can be thought of as concentrated sharply at $Pr(heads) = 1/2$, so that $Pr(Pr(heads) = 1/2) = 1$ (or some close approximation to that situation).[11] (Skyrms 1977, 707)

In the above passage, Skyrms claims that evidence E has changed the spread of John's 'second-order probabilities' over the spectrum for $Pr(H) = x$. However, Skyrms' usage of the term 'second-order probabilities' is, in my view, misleading. For what Skyrms calls 'second-order probabilities' is really John's subjective probability density distribution f of the *chance* of the coin landing heads $Ch(H)$. In other words, the difference between the ignorance situation and the ideal-evidence situation is that in the former John$_{\neg E}$ assigns a uniform probability density function to the chance of the coin landing heads $Ch(H)$, whereas in the latter John$_E$'s probability density distribution is concentrated sharply at $Ch(H) = 1/2$ so that $Pr(Ch(H) \approx 1/2) \approx 1$. Despite this difference, however, both John$_{\neg E}$ and John$_E$'s credence in H is $1/2$ since their *expectation* of the chance of the coin landing heads is $1/2$ in both cases (i.e. $Pr(H) = Pr(H|E) = \int_0^1 f(Ch(H) = x) \cdot x\, dx = 1/2$). Below is a picture to illustrate the difference between John$_E$'s and John$_{\neg E}$'s subjective probability density distributions f of the chance of the coin landing heads $Ch(H)$.

[11] For consistency, I have replaced all instances of 'tails' with 'heads'.

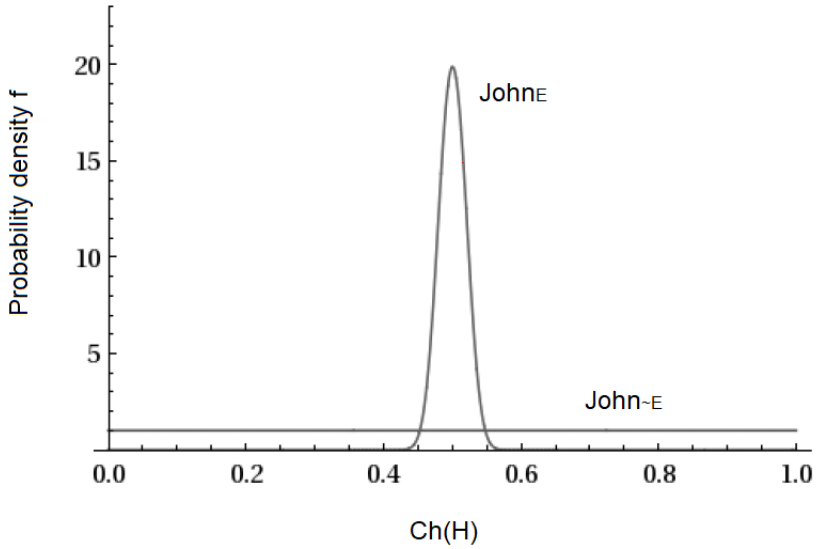

FIGURE 6.1: The effect of evidence E on John's probability density function f of the chance of the coin landing heads $Ch(H)$

So, informally, John$_E$'s credence in H is more resilient than John$_{\neg E}$'s relative to the observation of a long series of tosses all yielding heads because his probability density distribution is concentrated sharply at $Ch(H) = 1/2$ and hence it will take an extremely long series of tosses all yielding heads to shift this probability density distribution in such a way that its expectation of the chance of the coin landing heads is considerably different from $1/2$. Whereas the same cannot be said for John$_{\neg E}$.

The reason why it is important to stress that the claim that there is a difference between John$_{\neg E}$'s and John$_E$'s epistemic state relies on a distinction between chances and credences, rather than a distinction between probabilities and second-order probabilities is two fold. First, we can't invoke the notion of second-order probability without dramatically departing from the Bayesian framework. Hence if we interested in understanding the *Bayesian* response to Popper's concern, it seems to me that the only way to make sense of *why* the weight of evidence manifests itself in the resiliency of agent's credences in this example must be through a distinction between credences and chances. Second, all the examples on which the Bayesian relies to convince us that the weight

of the evidence manifests itself in the resiliency of an agent's credence in a hypothesis are always, to the best of my knowledge, examples in which an agent's credence in that hypothesis is mediated by her beliefs about the hypothesis's objective chances. But this is troubling: showing that the weight of evidence can manifest itself in the resiliency of an agent's credence in a hypothesis in cases where that credence is mediated by their beliefs about the hypothesis' objective chances is nowhere close to showing that the weight of evidence manifests itself in an agent's credence in a hypothesis in cases in which that credence is *not* mediated by their beliefs about objective chances. I will come back to this point shortly, but first it will be helpful to look at an other example in which the weight of evidence manifests itself in an agent's credences in a hypothesis.

In Popper's 'paradox of ideal evidence' $Ch(H)$ is a continuous variable. Hence, we had to introduce the notion of a subjective probability density distribution f of the chance of the coin landing heads Ch(H) to make sense of what was going on there. However, it is not hard to find simpler examples that Bayesians have used to argue that the weight of evidence is manifested in the resiliency of an agent's credence in a hypothesis H and in which $Ch(H)$ is a discrete, rather than a continuous variable. Here is one from Joyce (2005, 159):

> *Four Urns*: Jacob and Emily both start out knowing that the urn U was randomly chosen from a set of four urns $\{\text{urn}_0, \text{urn}_1, \text{urn}_2, \text{urn}_3\}$ where urn_i contains three balls, i of which are blue and $3 - i$ of which are green. Since the choice of U was random both subjects assign *equal* credence to the four hypotheses about its contents: $Pr(U = \text{urn}_i) = 1/4$. Moreover, both treat these hypotheses as statements about the *objective chance* of drawing a blue ball from U, so that knowledge of $U = \text{urn}_i$ 'screen offs' any sampling data in the sense that $Pr(B_{next}|E\&U = \text{urn}_i) = Pr(B_{next}|U = \text{urn}_i)$, where B_{next} says that the next ball drawn from the urn will be blue and E is a proposition that describes any prior series of random draws with replacement from U. Finally, Jacob and Emily regard random drawing with replacement as an exchangeable process, so that any series

> of draws that produces m blue balls and n green balls is as likely as any other such series, irrespective of order. Use B^mG^n to denote the generic event in which m blue balls and n green balls are drawn at random and with replacement from U. Against this backdrop of shared evidence, suppose Jacob sees five balls drawn at random and with replacement from U and observes that all are blue, so his evidence is B^5G^0. Emily, who sees Jacob's evidence, looks at fifteen additional draws of which twelve come up blue, so her evidence is $B^{17}G^3$. What should Emily and Jacob think about B_{next}?

In this example, Joyce clarifies that Jacob and Emily both treat the four hypotheses about the urn's contents '$U = urn_i$' as statements about the objective chance of drawing a blue ball from urn U. Hence, for clarity, I will replace the proposition '$U = urn_i$' with the proposition '$Ch(B_{next}) = i/3$' from now on. From the example, we can assume that Emily and Jacob assign *equal* credence to the four hypotheses about its contents, hence $Pr(Ch(B_{next}) = i/3) = \frac{1}{4}$ for all $i \in \{0, 1, 2, 3\}$ for both Emily and Jacob. And from David Lewis's principal principle, according to which an agent's credences should reflect objective chances (assuming the agent has no inadmissible information), we can also assume that $Pr(B_{next}|Ch(B_{next}) = i/3) = \frac{i}{3}$ for both Emily and Jacob. Hence we have all the information we need to apply Bayes' theorem to determine Jacob and Emily's posterior credences in B_{next}. John sees five blue balls and hence by applying Bayes theorem his posterior credences in the four chance hypothesis are as follows:

> Jacob:
>
> - $Pr(Ch(B_{next}) = 0|B^5G^0) = 0$
> - $Pr(Ch(B_{next}) = \frac{1}{3}|B^5G^0) = 0.0036$
> - $Pr(Ch(B_{next}) = \frac{2}{3}|B^5G^0) = 0.1159$
> - $Pr(Ch(B_{next}) = 1|B^5G^0) = 0.8804$

Jacob's expected value of the chance of the next ball drawn being blue is then $c(B_{next}|B^5G^0) = 0.959$. But Emily has seen fifteen additional draws of which

twelve came up blue, hence her posterior probabilities in the four chance hypotheses are as as follows:

Emily:

- $Pr(Ch(B_{next}) = 0|B^{17}G^3) = 0$
- $Pr(Ch(B_{next}) = \frac{1}{3}|B^{17}G^3) = 0.00006$
- $Pr(Ch(B_{next}) = \frac{2}{3}|B^{17}G^3) = 0.99994$
- $Pr(Ch(B_{next}) = 1|B^{17}G^3) = 0$

Emily's expected value of the chance of the next ball drawn being blue is then $Pr(B_{next}|B^5G^0) = 0.666626$.

Notice that Jacob has a much higher posterior credence in B_{next} than Emily. According Joyce this means that 'Jacob's total evidence points more decisively than Emily's does toward B_{next}'.[12] However, Emily has seen more draws than Jacob and hence has weightier evidence than him (regardless of whether we understand the weight of evidence as weight$_1$ or weight$_2$) and this fact is not reflected in Jacob and Emily's credence in B_{next} (because it can't be). However, as Joyce argues, the fact that Emily has weightier evidence than Jacob does manifest itself in the *resilience* of Jacob and Emily's credence in B_{next} relative to new evidence. For instance, 'if both subjects received evidence that tells against B_{next}, then Jacob's beliefs are likely to change more than Emily's. Suppose that both see five more balls drawn, and all are green. Jacob's credence will fall from near 0.96 to 0.5. Emily's will move hardly at all, dropping from 0.666626 to 0.666016' (Joyce 2015, 161). Indeed, Jacob's credence is 'less resilient than Emily's with respect to almost every potential data sequence, the sole exceptions being those sequences in which only blue balls are drawn' (ibid., 61), which is ultimately because Emily's credences are considerably less spread out over the possible chance hypotheses than Jacob's credences (e.g. notice that Emily has 0 or approximately 0 credence in three out of four chance hypotheses, whereas Jacob in

[12]To accept Joyce's claim that the fact that Jacob's posterior credence in B_{next} is higher than Emily's means that his 'total evidence points more decisively than Emily's does toward B_{next}' one must first and foremost accept that Bayesian inference is adequate for determining when one has evidence for a hypothesis. In this PhD, I have remained agnostic about whether this is the case, hence I am also agnostic about Joyce's claim.

two out of four). So, once again, this is a case in which the weight of evidence is manifested in the resiliency of an agent's credence in a hypothesis. And once again this is a case in which an agent's credence in the hypothesis are mediated through the agent's beliefs about objective chances.

So what to think of all this? Is the Bayesian's response to Popper's concerns satisfactory? There are at least two reasons for thinking that it is not. The first reason is that, as mentioned earlier, showing that the weight of evidence can manifest itself in the resiliency of an agent's credence in a hypothesis in cases where that credence is mediated by the agent's beliefs about objective chances is nowhere close to showing that the weight of evidence manifests itself in the resiliency of an agent's credence in a hypothesis in cases in which that credence is *not* mediated by their beliefs about objective chances.[13] In the above two examples, the reason for why weightier evidence tends to increase the resiliency of an agent's credence in a hypothesis has to do with the effect that the weight of evidence has on the agent's beliefs about the objective chance of that hypothesis (e.g. in both cases the weight of evidence reduces the number/size of the interval of chance hypotheses that the agents considers plausible). It is roughly due to *this* effect, that the agent's expected value of the chance of the hypothesis (i.e. the agent's credence in that hypothesis) tends to become more resilient as the weight of evidence increases. However, there is no analogous story to be told in cases in which an agent's credence in a hypothesis is not mediated through their beliefs about the objective chance of that hypothesis and hence the examples above do not give us any reason whatsoever for thinking that the weight of evidence manifests itself in the resiliency of an agent's credence in a hypothesis in those cases. And some very basic considerations show this idea to be rather dubious too.

Indeed, notice that in both of the above cases the agent's credence in the

[13]One may perhaps object: Popper's paradox of ideal evidence concerns a case in which the agent's credence is mediated by the agent's beliefs about objective chances. Hence, so the objection might go, from the outset the discussion has assumed that we're dealing with objective chances and so it's not clear why a change of setup would be an objection to the Bayesian's response to Popper's concerns. However, I see no reason for supposing that Popper's concerns should be restricted to examples in which the agent's credence is mediated by the agent's beliefs about objective chances. If Popper's concerns stem from the fact that the weight of evidence is not represented in an agent's credence, then a response to Popper's concerns which only applies to a restricted class of possible cases is clearly not a satisfactory one.

hypothesis tends to stabilize around the actual chance of the hypothesis. But in cases where there are no chances involved, that is when dealing with a hypothesis that has no objective chance of being true, the agent's credence in that hypothesis can't stabilize around the objective chance of the hypothesis since the hypothesis has no objective chance of being true! Hence what reasons could we possibly give to justify the idea the weight of evidence tends to stabilize the agent's credence in that hypothesis on any probability value that is neither 0 nor 1? It seems to me none. The hypothesis is either true or false and the best one can hope for is that as the weight of evidence increases the agent's credence in that hypothesis will tend to get closer and closer to 1 or to 0 depending on whether the hypothesis is true or false. But even this is clearly just a hope. As discussed in the previous section, regardless of whether we understand the weight of evidence as weight_1 or weight_2, the weight of evidence increases independently of the extent of the relevance of each incremental piece of evidence, that is independently of whether the additional evidence is highly in favour of H or highly in favour of $\neg H$. But then there is clearly no reason to suppose that as the weight of evidence increases an agent's credence in a hypothesis is bound to get closer and closer to 1 or 0 depending on weather the hypothesis is true or false.[14]

Could one perhaps argue that in cases where the agent's credence in a hypothesis is not mediated by objective chances, as the weight of evidence increases, an agent's credence in a hypothesis tends to become more resilient in a particular *range* of credences? For instance, if my credence in H is 0.9, then one might argue that the weight of evidence tends to increase the resiliency of my credence in the range $[0.9, 1]$. But if my credence in H is 0.2, then one might

[14]Savage (1972), and several others after him, attempted to show that under certain conditions a Bayesian agent's credence will converge to the truth with probability one. However, as many have pointed out (Glymour, 1980; Earman, 1992), these convergence to the truth results don't show that a Bayesian's credence will actually converge to the truth: all they show is that Bayesian agents are certain that they will, despite that not being necessarily the case. Hence these results are clearly too weak to underwrite a notion of objectivity and they may even 'constitute a real liability for Bayesianism by forbidding a reasonable epistemological modesty' (Belot, 2013). More recently Nielson (2020) has shown that for a Bayesian to be guaranteed to actually converge to the truth (rather than be certain that they are going to) the agent's priors must satisfy an extremely demanding condition (which he calls the strong regularity thesis), according to which 'there exists some positive real number that is strictly less than every probability of a non-empty proposition' (bid., 1463). This condition can only be satisfied by a finite probability space and hence is at odds with a substantial proportion of probability theory and its applications in statistics and the sciences.

argue that as the weight of evidence increases, my credence in H will tend to become more resilient in the range $[0, 0.2]$. To better understand this idea, consider the following example. Suppose that my credence in H: 'John stole my copy of Catch 22 in the library' is 0.9. I have high credence in H because e.g. I have lost my copy of Catch 22 in the library, I know that John stole another book from me in the past, and I recently saw him reading the same edition of Catch 22. However, I have yet to undergo a serious investigation of the relevant evidence (e.g. I still need to find out whether John was in the library when I lost the book, I need to talk to John's friends to find out whether they know if John had a copy of Catch 22 prior to me losing mine, I need to do a finger print test on the desk I was sitting at etc.). Suppose further that I do gather more relevant evidence and that my credence in H happens to remain 0.9. Could one perhaps argue in this case that although my prior credence in H was not very resilient in the range $[0.9, 1]$, since there was a lot of relevant evidence that I had yet to obtain which could have (substantially) decreased my credence in H (for instance, I could have found out that John was not in the library that day, or that he has finally decided to get around reading his copy of Catch 22, despite having bought it a couple years ago etc.), after having gathered more relevant evidence, and in particular after the weight of evidence has increased, my credence in H is now more resilient in the range $[0.9, 1]$?

This is, arguably, the only possible way to reconcile the idea that a greater weight of evidence is somehow correlated to the resiliency of an agent's credence in a hypothesis (in cases where the agent's credence is not mediated by objective chances). But, unfortunately, there are several problems with this idea. For a start, it is unclear how one would go about measuring the resiliency of a credence in a hypothesis (in a given range) when we have lots of different sorts of evidence. Should we perhaps think about the set of evidence that we could easily get, and see whether an agent's credence would move out of the range were they to get any of that evidence? If so, can one really circumscribe the set of evidence that one could easily get in any given case? If not, what set of evidence shall we consider? But even once we decide the relevant set of evidence with respect to which we should evaluate the resiliency of my credence in H,

how should we evaluate the overall resilience of my credence in H? Suppose, for instance, that my credence is very resilient with respect to some possible sorts of evidence but not resilient at all with respect to some other possible sorts of evidence. One may think that the overall resilience of my credence in H may perhaps depend on the ratio of the amount of evidence with respect to which my credence in H is resilient and the amount of evidence with respect to which my credence is not resilient. But is the evaluation of this ratio possible in light of the fact that, as discussed in Section 6.2, the principle of equipollance seems to be indefensible in most cases?[15] Secondly, regardless of whether we understand the weight of evidence as weight_1 or weight_2, the weight of evidence increases with each incremental piece of evidence independently of the extent of the relevance of each incremental piece of evidence and also independently of the relevance of the possible evidence that we have yet to consider. Hence, it seems that the weight of evidence can increase despite having very little if any effect on the resiliency of my credence in H with respect to the remaining evidence if the remaining evidence is highly relevant. So there doesn't seem to be a direct link between the weight of evidence and the resiliency of an agent's credence in a range.

To make this point more salient, suppose instead that my prior credence in H: 'John stole my copy of Catch 22 in the library' is 0.6. And suppose that despite having gathered lots of relevant evidence (e.g. some of which was highly in favour of H and some of which was highly in favour of $\neg H$), my posterior credence in H remains 0.6. Although the weight of evidence has increased, at the end of the day there are only two options: either John stole my copy or he didn't. Hence, despite having gathered a lot of relevant evidence already, there is no reason to suppose that if I were to gather even more relevant evidence my credence won't significantly depart from 0.6. And if so, why suppose that my posterior credence in H is now more likely to increase than decrease and hence

[15] In the coin tosses example the space of possible evidence is well defined (and also finite as long as we restrict our attention to a finite number of possible future coin tosses) and furthermore the principle of equipollance seems applicable in this case since it seems reasonable 'to suppose that the primitiveness or non-primitiveness of a predicate is unambiguously determined by the facts rather than convention' (i.e. whether I observed 2 rather than 3 coin tosses doesn't seem to be a matter of convention.)

be more resilient in the range $[0.6, 1]$ than my prior credence in H? The only reason for supposing this would be that in light of the extra relevant evidence that I have gathered, there are now more reasons to suppose that the remaining evidence is in favour of H than before I had gathered that evidence. However, there is nothing in the notion of the weight of evidence that remotely suggests this! Just for illustration, consider the following (bad) argument to justify why one should think that my credence in H is more resilient in light of the extra evidence I have gathered: the weightier the evidence, the more reasons for believing the hypothesis is true and hence the more reasons to expect that future evidence will be in favour of it. Recall, however, that the weight of evidence has nothing to do with how strongly one should believe a hypothesis. In this example, in particular, my credence in H is the *same* as before I had gathered the extra evidence. Hence by my own standards, even though the weight of evidence has increased, I do not think I have now more reasons for believing that H is true than I did prior to gathering that evidence. Therefore, this kind of reasoning to motivate why the weightier the evidence, the more resilient an agent's credence in a hypothesis will be in a range, is clearly not valid.

Overall, not only is it not clear how we should evaluate the resilience of a credence in a range in cases where we have lots of different possible sorts of evidence, but it is also not at all clear why we should think that, in cases where the agent's credence is not mediated by objective chances, the weight of evidence (whether we understand it as weight_1 or weight_2) is reflected in the resilience of a credence in a range (however we choose to measure it) in the first place.

The second reason for why the Bayesian's response is unsatisfactory is that the claim that the weight of evidence tends to increase the resiliency of our credence in a hypothesis is not particularly helpful to anyone who is interested in determining the extent of the weight of evidence or even simply comparing the weight of different bodies of evidence. But if the Bayesian is willing concede that the weight of evidence is just as important an aspect of the evidence as its balance then it seems like there is no reason why the Bayesian should only attempt to measure the latter and not the former. So this raises the following question: is it possible for the Bayesian to measure the weight of evidence? As far this

question is concerned, Joyce (2005) takes up the challenge.

Joyce (2005) argues that although the resilience of an agent's credences is often a reliable symptom of the weight of evidence as many before have observed, 'it is not the heart of the matter' (ibid., 166). Indeed, as seen in the examples above, the actual reason why weight tends to manifest itself in the resilience of an agent's credence in a hypothesis is because weightier evidence 'tends to cause credences to concentrate more and more heavily on increasingly smaller subsets of chance hypotheses, and this concentration tends to become more resilient' (ibid., 167). Hence, if this is what weight does, according to Joyce, we should try to measure the extent of *this* effect as an *indirect* way of measuring the weight of evidence. He proposes the following measure:[16]

$$w(H|E) = \sum_x |Pr(Ch(H) = x|E) \cdot (Pr(H) - x|E)^2 - Pr(Ch(H) = x) \cdot (Pr(H) - x)^2| \tag{6.2}$$

where E is some potential data proposition. According to him, the more an agent's credences are concentrated on a smaller subset of chance hypotheses and the more resilient this concentration is, the smaller the value of $w(H|E)$ will be. Hence, Joyce proposes to take $w(H|E)$ as an *indirect* measure of the weight of evidence, where the smaller the value $w(H|E)$ is, the weightier the evidence for H is supposed to be.

Although this is a welcome attempt to (indirectly) measure the weight of evidence, this measure has some problems and, importantly, some serious caveats too. I will start with the problems, some of which Joyce himself acknowledges. First, notice that the value of $w(H|E)$ crucially depends on the choice of E. That is, the value of $w(H|E)$ will be affected by what potential data proposition E we choose to consider. Joyce acknowledges this, but seems to suggest that we

[16] When x is a continuous variable, Joyce's measure would have to be the following:

$$w(H|E) = \int_0^1 |f(Ch(H) = x|E) \cdot (Pr(H) - x|E)^2 - f(Ch(H) = x) \cdot (Pr(H) - x)^2| dx, \tag{6.1}$$

where f is the density that defines the probability distribution Pr.

should not be too troubled, since $w(H, E)$ will be small for a wide range of potential data propositions E when the evidence is weighty and $w(H, E)$ will *not* be small for a wide range of E when the evidence is *not* weighty. However, regardless of whether this is the case, this doesn't change the fact that the value of $w(H|E)$ *will* be affected by the choice of E, and in some cases very much so. For instance, in the example above, Jacob's value of $w(H|E)$ will considerably be closer to Emily's value of $w(H|E)$ if say, we take E to be a sequence in which three blue balls are drawn, instead of a sequence in which two blue balls and one red ball are drawn.[17] So if we are to take this measure of weight seriously, then something must be said about how to choose E (or perhaps even a class of E). In light of this, this measure of the weight of evidence is at best incomplete. Second, notice that the value of $w(H|E)$ will invariably be affected by an agent's subjective priors in the different chance hypotheses. In the example above, the choice of the urn U is random and hence all agents must assign equal credence to the four hypotheses about its content. However, if that choice were *not* random, then different agents might assign different priors to the four hypotheses and this difference in priors would affect the value of $w(H|E)$. This means that distinct agents can obtain different values for $w(H|E)$ even if they have seen exactly the same evidence. Is this a problem? Well, it is if we want a measure of the weight of evidence that is independent of an agents prior beliefs, beliefs that have nothing to do with that evidence. But regardless of whether one is troubled by this, it is clear that the fact that the value of $w(H|E)$ depends on an agent's subjective priors, consists in an important departure from what Keynes had in in mind with his notion of the weight of evidence. Besides the fact that, as mentioned in the previous section, Keynes didn't think that we could have unconditional credences in the first place, Keynes's notion of the weight of evidence is completely independent from an agent's credences and Joyce's measure of the weight of evidence is evidently not; so there is a clear mismatch here. Third,

[17] In this example Jacob has been rather mislead by the evidence, so although his value of $w(H|E)$ will be very close to Emily's value of $w(H|E)$ for some choices of E, I don't think it will ever be smaller than Emily's value of $w(H|E)$ regardless of what potential data proposition E we choose. However, it should not be too hard to design an example in which Jacob is less mislead by the evidence, in such a way that his value of $w(H|E)$ will be smaller than that of Emily's for some choices of E, despite him having seen less balls than Emily.

many Bayesians (including Joyce), are not committed to the existence of sharp numerical degrees of belief. According to them a person's belief can be represented by a set of credence functions (Joyce calls this an agent's credal state) rather than just one credence function. Hence according to this view it is possible for an agent to assign more than one probability value to a hypothesis H. But then, if this measure of weight is to apply to imprecise probabilities, and not only precise probabilities, then something must be said about how we are supposed to calculate $w(H|E)$ in cases where there is more than one credal function in an agent's credal set, since this can give rise to a different value for $w(H|E)$ depending on what credal function one calculates its value relative to. Is one supposed to calculate $w(H|E)$ for *all* credence functions in an agent's credal set, and perhaps take an average of the different values we get? If not what shall one do? Regardless of what is the right answer to this question, it is clear that if we want to apply this measure of weight to imprecise probabilities as well as precise probabilities, we need one. So, this is an other reason why this (indirect) measure of weight is at best incomplete.

Leaving aside the problems mentioned above, this measure of weight has a crucial caveat: it is only applicable in cases where an agent's credences in a hypothesis are mediated by her credences about the objective chances of that hypothesis, and hence as Joyce himself acknowledges 'it's applicability is limited' (ibid., 166), which is, arguably, an understatement. This caveat entails that the idea that the Bayesian has finally found a comprehensive measure of weight is altogether unwarranted.

6.3.1 The weight of evidence and severity: two (very different) sides of the same coin?

According to a severe tester, one is justified in declaring to have evidence in support of a hypothesis just in case the hypothesis in question has passed a severe test, one that it would be very unlikely to pass so well if the hypothesis were false. Deborah Mayo calls this the *strong severity principle*:

Strong severity principle: We have evidence for a claim C just to

> the extent it survives a stringent scrutiny. If C passes a test that was highly capable of finding flaws or discrepancies from C, and yet none or few are found, then the passing result, x, is evidence for C. (Mayo 2018, 14)

In her extensive and persuasive defence of statistical inference as severe testing (or error statistics), Mayo (2018) argues that Bayesian inference is unable to guarantee that the above principle will be met: a Bayesian can declare to have evidence for a claim despite not having done anything to severely test that claim, since the posterior probability of a hypothesis does not directly depend on the severity of a test it has passed. The core reason for this has to do the (infamous) *likelihood principle* whose violation is not an option for anyone who subscribes to the Bayesian paradigm. Here is a statement of it by Berger and Wolpert (1988, 19):

> **The likelihood principle**: All the information about θ obtainable from an experiment is contained in the likelihood function for θ given x. Two likelihood functions for θ (from the same or different experiments) contain the same information about θ if they are proportional to one another.

However, for anyone who believes in the strong severity principle, the likelihood principle must be wrong since it entails that inferences about hypotheses depend exclusively on the outcome of an experiment, and not on its design. But considerations pertaining to the design of an experiment (e.g. whether someone deliberately stops an investigation depending on what the data looks like) are relevant for assessing the severity with which a hypothesis has passed a test, and consequently whether one has evidence in support of a hypothesis.[18]

The Bayesian, however, seems to be largely unmoved by the incompatibility between the strong severity principle and the likelihood principle. According to the Bayesian all one needs to obtain posterior probabilities in various hypotheses is the prior probabilities of those hypotheses and their likelihood function given

[18]Disagreement about the likelihood principle is a core issue in the philosophical debate between frequentists and Bayesians.

the observed data, 'consequently the whole of the information contained in the observations that is relevant to the posterior probabilities of different hypotheses is summed up in the values that they give to the likelihood' (Jeffreys 1961, 57). Hence, the likelihood principle must be right and the strong severity principle wrong, if incompatible with it.[19]

The reason I mention all this is not to argue whether or not one should pick the strong severity principle over the likelihood principle. Rather, I mention this because, despite the fact that the notion of severity has very little to do with that of the weight of evidence,[20] I think the Bayesian's never ending quest for some way to account for the latter betrays the Bayesian's confidence in the likelihood principle after all. That is, the Bayesian's recognition that the posterior probability that one assigns to a hypothesis given the available evidence (in light of the agent's priors and likelihood function) is unable to reflect something that seems important about the nature of that evidence (i.e. its weight) is, in my view, a mere symptom of the Bayesian's *own* dissatisfaction with the likelihood principle: there is more to say about the evidence for hypotheses than their likelihood function given that evidence and the Bayesian can't account for what that something is.

As argued in Section 6.3, although there have been various attempts to account for the weight of evidence, none of them can show that the Bayesian can account for it in cases where an agent's credence in a hypothesis is not mediated

[19] Van Dongen et al. (2020) have recently argued that the Bayesian can cash out severity in terms of the expected success of the predictions a theory makes with respect to its negation. Hence, according to them what it takes for a test to be severe is that the tested hypothesis imposes 'substantial restrictions on the range of potential data that are consistent with it' (ibid., 13). They further argue that 'Popper and Bayes can thus be reconciled: the evaluation of hypotheses in terms of Bayes factors is influenced by their specificity and Bayesian inference has the conceptual resources to reward specific predictions' (ibid., 21). However, the severity of a test according to this account is unrelated to the design of the experiment (e.g. whether or not a test is severe according to this account is unaffected by whether someone deliberately stops an investigation depending on what the data looks like) and hence is, in my view, an unsatisfactory account of severity.

[20] Why do I claim that the notion of severity has very little to do with the notion of the weight of evidence? To see this, consider weight_1 for simplicity. Weight_1 will increase as more data comes in independently of whether someone deliberately stops an investigation based on what the data looks like. That is weight_1 will increase as the data increases, regardless of how that data was acquired (i.e. regardless of the experimenter's intentions). But to the severe tester what matters in the assessment of whether one has evidence in support of a hypothesis is not how much relevant evidence one has gathered, rather it is whether or not the process of generating that evidence has been able to severely test the hypothesis in question. Similar considerations apply to weight_2. Hence the weight of evidence, however we choose to understand it, seems to have very little to do with the notion of severity.

by objective chances. Furthermore, as discussed in Section 6.2, any attempt to directly measure the weight of evidence also seems impossible. But then what to make of this? Should the Bayesian's worries about 'the weight of evidence', a concept that has proved to be extremely hard (and that is arguably impossible) to measure either directly or indirectly, a reason to try harder or perhaps a reason to abandon the Bayesian framework all together? Although I merely wish to raise the question, without answering it, it is worth pointing out that under a severe testing perspective the problem of the weight of evidence becomes some sort of a pseudo problem. For the severe tester what matters in the assessment of whether one has evidence in support of a hypothesis is whether or not the process of generating that evidence has been able to severely test the hypothesis in question; hence whether one has 'a lot' or 'a little' evidence without an understanding of the process that generated that evidence is irrelevant. Perhaps then the weight of evidence and severity may be thought of as two (very different) sides of the same coin: they are two unrelated notions, but what brings them together is the fact that they both seem to make trouble for the likelihood principle, a principle at the core of Bayesian inference.

Chapter 7

An assessment of some proposals for a new IPCC uncertainty framework

7.1 Introduction

Throughout this chapter, I will assume that an adequate uncertainty framework for the IPCC should (at the very least) satisfy the following two desiderata:

1. the framework's fundamental concepts (e.g. probability, likelihood, confidence etc.) should be clearly defined so that they can be used appropriately and consistently by the IPCC authors in the communication of uncertainty;

2. the use of the framework's fundamental concepts should help the IPCC authors produce findings that are interpretable, relevant and useful for the target audience/s (e.g. policy makers, decision makers, the general public).

In Part 1, I have argued extensively that the current IPCC uncertainty framework fails to adequately meet the above desiderata, concluding that a better and more carefully considered framework is in urgent need. The aim of this section is to critically assess the extent to which some recent proposals for a new IPCC uncertainty framework are an improvement over the current one with respect to the above desiderata. I won't refrain from criticizing, when I believe there

are criticisms to be made. However, I will try my best to identify the merits of each proposal and to offer some constructive guidance for a better future IPCC uncertainty framework.

The structure of this chapter is as follows. In Section 7.2, I will critically assess Winsberg's (2018) proposal, according to which the likelihood metric should be used to communicate the range of credences that the IPCC authors assign to a hypothesis in light of the available evidence, and the confidence metric should be used to communicate 'how likely their consensus regarding appropriate credences is going to remain fixed in the light of future developments' (ibid., 105). Amongst other things, I will argue that Winsberg's interpretation of the confidence metric, under his own proposal, is unjustified. In Section 7.3, I will assess Mach et al.'s (2017) proposal, which gets rid of the confidence metric and replaces it with qualitative terms for scientific understanding. I will argue that Mach et al.'s proposal faces very similar conceptual problems to the current uncertainty framework (problems which were discussed extensively in Part 1) and that it is, therefore, not a considerable improvement over the current uncertainty framework with respect to the above desiderata. In Section 7.4, I will identify some of the merits of both Winsberg's and Mach et al.'s proposal and I will incorporate them into my own tentative sketch of a proposal for a better IPCC uncertainty framework. Finally, in Section 7.5, I will assess Bradley et al.'s (2017) proposal for a new IPCC uncertainty framework, which stems from a desire to help clarify what role probability ranges, qualified by confidence judgments, should play in decision making. Despite the more than justified motivation behind this proposal, I will argue that the interpretation of confidence under this proposal is problematic. Hence, this proposal fails to meet the first desideratum. I will further argue that my own proposal does clarify how the IPCC findings should be interpreted by decision makers and hence addresses Bradley et al.'s concerns without having to rely on an overly problematic interpretation of confidence.

7.2 Winsberg's proposal: Can scientists measure the resilience of their credences, and should they?

As briefly mentioned in Chapter 1, Winsberg (2018) proposes a particular interpretation of the likelihood metric and the confidence metric in the IPCC uncertainty framework. In essence, according to his proposal the likelihood metric is supposed to communicate the range of credences that the IPCC authors assign to a hypothesis in light of the available evidence. The confidence metric, on the other hand, should be used to communicate 'how likely their consensus regarding appropriate credences is going to remain fixed in the light of future developments' (ibid., 105). Hence according to Winsberg, confidence can be thought of 'as a kind of second-order probability – since, in effect, it would reflect the panel's estimate of the likelihood of their credence changing in the future' (ibid., 105). In Section 1.4.1, I have argued that Winsberg's interpretation of likelihood and confidence is incompatible with some of the Guides' recommendations- and thus with the resulting practice of the IPCC authors in their communication of uncertainty. Hence, if Winsberg's proposal is to be implemented coherently, some of the guide's recommendations would evidently have to change. The aim of this section, however, is to go deeper into Winsberg's proposal and assess whether it is a good one in the first place.

Likelihood

According to Winsberg's proposal the likelihood metric should be used to communicate the 'range of probabilities that is satisfactory to almost all, if not all, of the members of a panel. Group credence, in other words, should be considered to be the number, or range of numbers, that is the consensus of the group regarding what one's degree of belief in the hypothesis ought to be' (ibid., 103). There is a caveat, however. According to Winsberg, we should interpret credences not as things that scientists simply have but rather as 'things that scientists accept':

> There are of course situations in which a scientist might realize that the degrees of belief that they happen to have do not reflect the best

> available evidence. Or they might just not have degrees of belief in the purist's sense: they might have no particular disposition to bet on climate hypotheses. Even worse, if probabilities are just things that you have, it is difficult to see how a group can deliberate together about what the correct ones are. Following Cohen (1995) and Steel (2015), therefore, I think we should think of personal probabilities as things that scientists accept. On this view, the careful scientist will use whatever methods of reasoning are at her disposal to correct the degrees of belief that she happens to have, and in the case of the model based science we are discussing, she will attempt to estimate what the best probabilities for her to use are, based on her best models. And she will use Bayes' rule as a way of calculating probabilities, rather than as a constraint on rationality. (Winsberg 2018, 114)

In a nutshell, the idea here is that a subjective probabilistic assessment of the available evidence must rely on a probability model that specifies a likelihood function and a prior probability and that scientists might often have to make a choice about what model(s) to accept (and reject!), which will in turn determine what posterior probability they accept to assign to a hypothesis in light of the available evidence. Winsberg's view of credences as things that scientists often *accept* rather than simply have is reasonable.[1] However, if this idea is right, then it clearly raises the question as to what determines the decision to accept a probabilistic model over another. That is, if a scientist must decide to accept a probability model among multiple probability models all compatible with their beliefs, what determines the scientist's decision? In particular, if we take the idea of the acceptance of credences seriously then, as argued by Steel (2015), it seems like any such acceptance will be subject to an argument from inductive

[1] And I would also be willing to embrace the stronger claim that *any* Bayesian assessment of the available evidence in science *always* involves a prior *decision* about what probabilistic model to *accept*. In other words, I am happy to deny that precise subjective probabilities/credences can ever be plausibly thought as representations of a scientist's involuntary cognitive state. However, a defence of this stronger claim is beyond the scope of this thesis.

risk (Rudner, 1953).[2] That is, if scientists are in the business of accepting and rejecting (subjective) probabilities to assign to a hypothesis and these decisions can have implications for practical action then it seems that these decisions should depend in part on non-epistemic value judgments about the costs of error.[3]

But even if one agrees with the idea that there might often be more than one probability model to accept, perhaps one could question whether scientists are in fact forced to make a choice about which one to accept in order to report their beliefs, and hence question the idea that in such cases the reporting of scientists' beliefs must inevitably be subject to an argument from inductive risk. Indeed, one might suggest that the fact that the IPCC authors can report a *range* of probabilities (by assigning a particular likelihood level) rather than precise probabilities they can perhaps avoid decisions about which probabilities to accept. This suggestion effectively relies on the idea that IPCC authors can avoid making a decision about which probability models to accept by representing vague and incomplete degrees of belief by a complete set of probability functions and then determining all the probabilities that are rational to assign to a hypothesis according to these functions.[4] However, there are several problems with this suggestion. For a start, suppose one were to accept this as

[2]Rudner (1953) famously argued that given that the evidence is never enough to establish a hypothesis with certainty, scientists are always faced with a decision as to whether that evidence is sufficiently strong to accept or reject a hypothesis H. And ethical values (e.g. how bad real-world consequences would be were one to mistakenly accept or reject H) should affect this decision. Jeffrey (1956) famously challenged this argument by denying that scientists accept or reject hypotheses in the first place; according to Jeffrey scientists merely assign probabilities to hypotheses. However, if as Winsberg argues 'we should think of personal probabilities as things that scientists accept' rather than representations of scientists' involuntary cognitive state, then the decision of a scientist to accept a credence seems once again to be subject to an argument from inductive risk.

[3]Winsberg (2012, 2018) also argues that non-epistemic values can affect the probability values that climate scientists assigns to a hypothesis. However, he argues this on the basis that in climate science, 'scientists' best attempts at estimating $Pr(H|e\&\mathbf{B})$ [where e is a new piece of evidence and **B** is the scientists' background knowledge] will often involve estimating $Pr(H|e\&\mathbf{B'})$ instead, where **B'** replaces some of the claims in **B**, with a computationally tractable scientific model, M or set of models **M**, of the system or phenomenon under investigation [...] the distortions relative to **B** that scientists are willing to tolerate when developing **M**, however, will depend in part on the purposes and priorities of their investigations, as well as the purposes and priorities that shaped any earlier layers of the model's development' (Winsberg 2018, 146). But whether or not Winsberg's argument is successful in showing that social values affect what probabilities climate scientists assign to a hypothesis (see e.g. Parker (2014) for some possible objections) is irrelevant to the idea that if probabilities really are things that scientists accept then this acceptance is subject to an argument from inductive risk.

[4]This is a popular suggestion. For instance, Steele (2012) argues that 'scientists can simply report their beliefs to policy makers, using whatever representation that best "captures" these beliefs, whether this be a probability function, a set of probability functions, nonadditive probabilities, or something else' and hence do not have to 'choose their beliefs in a manner that takes into account real-world consequences' (897).

a reasonable suggestion and hence accept that 'the push-the-problem-one-step-back argument [from inductive risk] is misguided or at least incomplete' (Steele 2012, 897) since scientists can in principle report their beliefs using whatever representation best captures their beliefs. This on its own is not enough to show that scientists (qua policy advisers) can always avoid making value judgments when reporting their beliefs. Since as Steele (2012) argues, whenever scientists (qua policy advisors) must convert their beliefs to a standard scale (such us, for instance, the likelihood metric in the IPCC uncertainty framework), 'scientists cannot avoid making value judgments, at least implicitly, when deciding how to match their beliefs to the required scale' (Steele 2012, 899). Second, there are good reasons for not thinking that this suggestion is a reasonable one in the first place. For as Steel (2015) remarks, there are at least two difficulties with it:

> First, it does not avoid the problem of vagueness. For not only may a person's exact degrees of belief be vague, it may also be vague which probability distributions are consistent with her degrees of belief and which are not (Howson & Urbach 1993, 88-89). Thus, decisions would have to be made about which ensemble of personal probability models to accept. Secondly, this approach has the potential to greatly increase the complexity of probabilistic reasoning (Howson & Urbach, 1993). Instead of one possibly already quite complex probability model, one must consider a massive and potentially ill-defined set of models. As a result, the approach of representing degrees of beliefs by means of sets of probability functions comes with a practical cost of increased mathematical and computational complexity. (Steel 2015, 6)

But if the set of *all* probability models that are compatible (and only those) with a scientist's beliefs does not seem to be something that a scientist is able to consider in practice, as Steel suggests in the above passage, then the fact that the IPCC authors can report a range of credences in a hypothesis H does not mean that they can avoid decisions about which probabilities to accept. Why? For two reasons. First, if the IPCC authors don't know the probabilities that all

those probabilistic models compatible with their beliefs give to a hypothesis H, then neither can they know that all those probabilities should lie in the range of probabilities reported. Hence, a decision would still have to be made as to what range of probabilities to accept such that it includes all probabilities that all those unconsidered probabilistic models give to H. Second, and crucially, accepting that it is rational to assign a probability to a hypothesis in light of the available evidence (without there being a probabilistic model that is compatible with it and the scientist's beliefs) is no less problematic than not accepting it. That is, if a probability value for a hypothesis is not entailed by any probabilistic model that is compatible with the scientists' beliefs then that probability value should not be included in the reported range. Suppose, for instance, that one were to increase the range of probabilities reported in H without a worry in this world. In this case, one would effectively be willing to accept that it is rational to hold a particular credence in H in light of the available evidence, despite not having any good reason for accepting it. This last point is, in my view, particularly important, especially in relation to the IPCC current practice in their treatment of uncertainties. As discussed in Section 1.5 and Section 2.5, the IPCC authors often assign a *likely* level to a finding (a rather wide probability range i.e. (0.66, 1)) to account 'for additional uncertainties or different levels of confidence in models' (IPCC 2013, 23). However, the rationale behind how they arrive at this interval is not explained. Underling this practice is, in my view, the idea that claiming that the probability of an event is in the interval (0.66, 1) is less strong than claiming the probability of that the event is in a smaller interval e.g. (0.9, 1). This idea would make sense if we were talking about *objective* probabilities. In this case the claim that the objective probability of an event is in the interval (0.66, 1) would evidently be less strong than the claim that it lies in the interval (0.9, 1) since the former probability interval both includes and is greater than the latter. However, under the view that the reported range of probabilities is supposed to represent all and only those credences that the IPCC authors think it is rational to assign to a hypothesis in light of the available evidence, then the claim that it is rational to assign any credence in the range (0.66, 1) to a hypothesis is no less strong that the claim that it is only rational to assign credences

in the range (0.9, 1); since claiming that it is rational to assign a credence to a hypothesis in the interval (0.66, 0.9] is no less strong than the claim that it is *not* rational to do so. Hence, if likelihood is really to be understood as the range of all and only those credences that the IPCC authors accept it is rational to hold in a hypothesis in light of the available evidence, as Winsberg proposes, then the IPCC authors under this proposal should, arguably, not take the decision to accept probabilities as light-heartedly as the current practice suggests.[5]

Confidence

The interpretation of the likelihood metric under Winsberg's proposal is relatively clear, so it is time to turn to the confidence metric. How is "confidence" evaluated and what type of uncertainty is it supposed to represent under Winsberg's proposal?

Winsberg (2018, 105) suggests that there are three factors that a policy maker (or any other decision maker) might want to know in addition to the range of credences that an IPCC panel assigns to a hypothesis in light of the available evidence:

- 'how many different sources of evidence were consulted by the experts in arriving at the assessments of probability';
- 'how univocal (or the contrary) those various sources were';
- 'the degree to which the reported consensus of the committee papered over internal disagreement or, to the contrary, reflected easy-to-come-by agreement'.

According to Winsberg:

> All three of these factors are, in principle, independent. There could be many sources of evidence that disagree, or few sources that agree.

[5]Steel's argument for why the acceptance of probabilities in light of the available evidence is subject to an argument from inductive risk relies on the idea that the experts must accept a probabilistic model (or a set of them) to begin with. However, as discussed in Section 2.5.2, the IPCC does not seem to currently rely on Bayesian inference to determine the range of probabilities to assign to a hypothesis in light of multi-model ensemble's results. So the question of why they accept those probabilities in the first place is, in my view, a much more pressing question than whether this acceptance may be subject to an argument from inductive risk.

> There could be agreement or disagreement among the experts either way. Thus, in principle the policy maker might want to use all of this information as a guide to how she should act. She might use that added information, in conjunction with the reported probabilities, in a variety of ways. More conveniently, though, the policy maker might want the committee to summarize these three components of information into one single metric [i.e. the confidence metric]. (Winsberg 2018, 105)

Hence, under Winsberg's proposal, the role of the confidence metric is to summarize the three factors above into a single metric. I will discuss in some more detail each of those factors shortly, but first: why is, according to Winsberg, a summary of the evaluation of these three factors useful to policy makers? This is what he says:

> One way to think about this kind of self-assessment of confidence is as the committee's assessment of the degree to which the answers to the above questions foretell a resiliency in their credences; as an assessment of how likely their consensus regarding appropriate credences is going to remain fixed in the light of future developments (be they in modeling, physical understanding, data acquisition, etc.). One possible way to understand measures of confidence, therefore, might be as a kind of second-order probability – since, in effect, it would reflect the panel's estimate of the likelihood of their credence changing in the future. A high confidence in a credence is a bit like a high degree of belief that that credence will be resilient in the face of future evidence – assessed by looking at the variety of evidence supporting the credence, and the degree of agreement among those sources and among experts. But given the general murkiness of second- order probabilities in general, the lack of an obvious set of decision rules to apply to them, and the difficulties that would be involved in interpreting such probabilities in this specific case, I'm inclined to think that it is wise of the IPCC to refrain from using the

> expression "probability" for its second- order characterizations, and to limit itself to qualitative characterizations of confidence. (Winsberg 2018, 104)

So according to Winsberg, knowing the level of confidence, in addition to the range of credences that the IPCC experts assign to a particular hypothesis (i.e. in addition to the likelihood level), is useful for policy makers because the confidence level (under Winsberg's proposal) tells them the extent to which the IPCC experts believe that 'their consensus regarding appropriate credences is going to remain fixed in light of future developments'. So what Winsberg is essentially arguing is that, if the role of confidence is that of summarizing those three factors mentioned above, it is reasonable to suppose that the higher the confidence level is, the more strongly the IPCC experts believe that their credences in a hypothesis will remain fixed in light of future evidence. So under this view, if the experts claim to have a *high confidence* that the Equilibrium climate sensitivity (ESC) is *likely* in the range [1.5°C, 4.5°C]' then what this should mean, under Winsberg's proposal, is that the experts strongly believe that they will still consider this range for ECS to be *likely* in light of future evidence (as opposed to e.g. *unlikely*, *more likely than not* or *very likely*).

I will argue that Winsberg's interpretation of what the confidence metric is supposed to represent under his own proposal cannot be correct. But first it will be helpful to think a little about how to interpret the three factors on which the evaluation of the confidence metric depends under Winsberg's proposal.

Three factors relevant to the evaluation of confidence

The first factor Winsberg mentions is 'how many different sources of evidence were consulted by the experts in arriving at the assessments of probability'. But how should we interpret this factor? Could this be roughly what Keynes had in mind with his notion of 'the weight of evidence' and what he himself and the Bayesians have been struggling to represent/measure ever since? And if so, is it supposed to be a measure of the absolute amount of available evidence (corresponding to something like weight_1) or rather is it supposed to be a measure of the extent to which the available evidence is complete (corresponding to

something like weight$_2$)? Or perhaps is Winsberg thinking of something else all together?

To get a glimpse of how Winsberg might be interpreting this factor, it will be helpful to consider his discussion of the various different sources of evidence underlying the following IPCC finding:

> Equilibrium climate sensitivity (ESC) is *likely* in the range 1.5°C to 4.5°C with *high confidence*. ECS is positive, *extremely unlikely* less than 1°C (*high confidence*) and *very unlikely* greater than 6°C (*medium confidence*).

Winsberg argues that Schupbach's account of ERA diversity (discussed extensively in Part 2) can help us understand 'why, when it comes to the claim that it is "very unlikely" that ECS > 6 °C the IPCC only claims "medium confidence"'(Winsberg 2018, 206). Why is that?

Winsberg begins by pointing out that there are various rather different sources of evidence that are relevant for assessing the value of ECS. An important source of evidence, for instance, comes from the range of values for ECS predicted by multi-model ensembles. An other comes from observations of the post-industrial warming of the ocean and atmosphere in response to various external forcings (e.g. increasing concentrations of greenhouse gases, aerosols, volcanic eruptions etc.). Yet another one comes from paleoclimate records (such as the cooling of the Last Glacial Maximum or the last few glacial cycles). Winsberg further points out that the various methods that are used to estimate the correct value for ECS based on each of these different sources of evidence are subject to different sources of uncertainty. For instance, methods which rely on Paleoclimate records are particularly subject to measurement uncertainty since both the reconstructed past climate and forcing 'are inferred from indirect evidence that may not be spatially representative or may be responding to multiple factors, uncertainties that are difficult to quantify' (Knutti et al. 2017, 729). Measurement uncertainty is less salient when it comes to methods that rely on more recent observations of the climate response to forcing. However, these methods are particularly subject to uncertainty concerning whether of not the planet

has had time to reach equilibrium before the forcing was taken away (and so the worry is that one might be observing a transient response instead). And of course methods that rely on multi-model ensembles' results are subject to uncertainty as to whether those models adequately represent the climate system in spite of possible discrimination errors, parametarization uncertainty, the omission of important sources of feedback etc.[6] According to Winsberg, by thinking of each of these different sources of uncertainty as possible alternative explanations for the range of ECS values obtained by each of these methods we can inquire about the extent to which these methods are ERA diverse with respect to a target explanation (e.g. that the ECS value actually lies within a particular range) and its competitors:

> If we are interested in doing robustness analysis on these various detection methods, then it is helpful to think of each of these sources of uncertainty [...] as alternative possible explanations of various hypotheses detections. Suppose, for example, that using instrument data associated with a particular volcanic eruption, we find that the data support the hypothesis that ECS is between 1.5°C and 4°C. We can count this as a method of detection for this hypothesis. Thus, to do RA, we would want to ask: in addition to the truth of the hypothesis, what other explanations are there for the fact that this method detects that hypothesis? (Winsberg 2018, 205)

Once we have answered this question we can, according to Winsberg, proceed to consider other methods of estimating ECS that are able to rule out these alternative rival explanations. For instance, an alternative explanation for why we obtained this range for ECS using data associated with a particular volcanic eruption might be that the the climate hadn't yet reached equilibrium before the forcing was taken away and hence under this rival explanation what we observed was merely a transient response. Given that Paleoclimate data is less

[6] There are many other important sources of uncertainties when it comes to each of these three methods for estimating the ECS, many of which are mentioned by Winsberg (2018, 203-204). But for the purpose of understanding Winsberg's analysis, this will suffice (see also Knutti et al. (2017) for a comprehensive summary of the current evidence relevant to the estimation of the ECS).

susceptible to uncertainty about whether the planet has reached equilibrium before the forcing was taken away (as long as the actual value of ECS is not very high since the larger the value of ECS is, the more slowly equilibrium is expected to be reached) we can, according to Winsberg, use methods that rely on Paleoclimate data to rule out this explanation for the detection that ECS > 1.5°C. And by considering more and more detection methods, we can, rule our more and more alternative explanations for why detect that ECS > 1.5°C.

In contrast, Winsberg argues that it is more difficult to find detection methods that are able to rule out alternative explanations for why we detect that ECS < 6°C:

> What is of course interesting is that this set of detection methods is very good at ruling out alternative explanations for the hypothesis that ECS > 1.5 °C, but not very good at all at ruling out alternative explanations of the hypothesis that ECS < 6 °C. One good way to see this is to ask: if ECS were greater than 6°C, what would explain all of our detections that it is lower? Unfortunately, it is not hard to come up with explanations: suppose, for example, there is a strong but as-yet-unaccounted-for positive feedback mechanism. Then we would not expect our models to correctly detect the high value of ECS, and we would not expect our instrument records to detect it either, because (being a high value of ECS) it would act too slowly for them to see it. We would probably only expect to see it in the millions of-years-scale paleodata – but those data sets have enough uncertainty that they are poor at eliminating such a hypothesis [...] (Winsberg 2018, 205-206)

Winsberg concludes that this is the reason why the IPCC authors assign only a "medium confidence" to the claim that it is "very unlikely" that ECS > 6 °C and "high confidence" to the other probabilistic claims in the IPCC finding above:

> Robustness analysis helps us to see why we have high confidence that ECS is greater than 1.5°C, but lower confidence that it is less than 6°C, and virtually none at all regarding any hypothesis that is

> more fine-grained than 1.5°C < ECS < 4°C. [. . .] we can also see why the IPCC finds it useful to put probabilities on these hypotheses combined with a further estimate of confidence. Climate scientists seem to believe that each detection method puts a low probability on ECS < 1.5°C and on ECS > 6°C. Both are statistical outliers in each of the methods. But the RA we just performed on each hypothesis reveals that the ECS < 1.5°C probability estimate is likely to be much more resilient – precisely because it is more robust. This is presumably why, when it comes to the claim that it is "very unlikely" that ECS > 6°C, the IPCC claims only "medium confidence." (Winsberg 2018, 206)

Hence, in a nutshell, according to Winsberg the confidence level that is assigned to a probabilistic claim concerning a particular hypothesis is affected by the extent to which the available evidence is ERA diverse with respect to that hypothesis and its competitors.

There are at least a couple of problems with Winsberg's analysis. The first is that, according to Schupbach's account of ERA diversity, the extent to which the detection methods are ERA diverse with respect to a target hypothesis (and its competitors) should affect the extent to which the target hypothesis is confirmed. In particular, if the detection methods are not ERA diverse with respect to the hypothesis ECS < 6°C, as Winsberg argues above, then according to Schupbach's account of ERA diversity, once's credence in that hypothesis should not increase. But then under Winsberg's analysis, it is unclear on what basis the authors are justified in declaring that it is "very unlikely" that ECS > 6°C.

A second crucial problem is that, contrary to what Winsberg suggests, it is hard to see how any of the methods he mentions for estimating the value of ECS are able to rule out competing hypotheses with respect to any target hypothesis in the sense required by Schupbach's account of ERA diversity. For instance, the fact that methods that rely on Paleoclimate data agree with methods that rely on instrument data associated with a particular volcanic eruption in so far as ECS > 1.5°C is not able to rule out that the latter result was due a transient response

to forcing. As mentioned earlier, Paleoclimate data is subject to a great deal of measurement uncertainty and hence it is more than possible that the these two methods could agree that ECS > 1.5°C, despite the former result being due to a transient response and the latter being due to a measurement error. Given this, it seems to me that Winsberg is effectively relying on some sort of argument from coincidence in the above analysis and not, as he suggests, on Schupbach's account of ERA diversity.

Leaving aside these problems with Winsberg's analysis, it seems that according to Winsberg the factor 'how many different sources of evidence were consulted by the experts in arriving at the assessments of probability' may, perhaps, be better interpreted (at least as far as this example is concerned) as something like 'how many methods that are subject to different types of uncertainty are used by the IPCC experts to arrive at the assessments of the probability of a hypothesis'. In any case, it seems clear that what the first factor is supposed to represent and how one should evaluate it is left rather open to interpretation.

The second factor that Winsberg argues should affect the evaluation of confidence underlying a probabilistic statement is 'how unequivocal (or contrary) various sources of evidence were'. If the level of confidence is affected by how unequivocal the sources of evidence are then confidence, under Winsberg's proposal, is clearly not anything like Keynes's notion of the weight of evidence. But, crucially, a probabilistic assessment of the evidence with respect to a hypothesis should surely be affected by the extent to which the various sources of evidence are unequivocal or contrary with respect to that hypothesis. Hence, if this factor plays a role in the evaluation of confidence, then confidence cannot be reasonably thought of as an independent dimension from likelihood.

The final factor that is supposed to be relevant for the evaluation of confidence according to Winsberg is 'the degree to which the reported consensus of the committee papered over internal disagreement or, to the contrary, reflected easy- to- come-by agreement'. This is no longer an evaluation of the evidence itself, but rather it is an evaluation of the extent to which the relevant experts agree on the range of probabilistic values that should be assigned to a hypothesis in light of the available evidence. What is particularly odd about this factor

is that it seems to be irrelevant, if we understand likelihood as Winsberg told us we should understand it. That is, if according to Winsberg the likelihood metric should be used to communicate the 'range of probabilities that is satisfactory to almost all, if not all, of the members of a panel', then by definition there should be very little if any disagreement regarding the range of probabilities that are ultimately assigned to a hypothesis (i.e. the assigned likelihood level). What ads an extra layer of confusion to this is that, according to Winsberg, these three factors are all in principle independent. As far this factor is concerned, this means that disagreement amongst the experts is in principle independent of how many different sources of evidence were consulted by the experts and how unequivocal (or contrary) those various sources were. From a Bayesian perspective, the only way to make sense of this seems to be if disagreement amongst experts as to what is the range of satisfactory probabilities that should be assigned to a hypothesis in light of the evidence is merely due to a disagreement as to what priors are reasonable. But this is just one source out of the many possible sources of disagreement amongst experts. As Steel (2015) remarks 'probabilistic assessments of evidence or degrees of confirmation depend on accepting data, background knowledge, and probability models' and whether or not to accept data or a particular probability model is directly affected by what sources of evidence the experts considers to be relevant for the hypothesis in question and also the extent to which the experts thinks the evidence is unequivocal or contradictory. So it seems more natural to think that disagreement amongst experts might more often than not stem from disagreement as to the evaluation of the first two factors, raising the question as to what role disagreement among experts plays in their evaluation.

Winsberg's interpretation of confidence

Leaving behind the details of how these factors should be evaluated and the questions that each of them raises, recall that according to Winsberg, if the confidence metric is a sort of summary of the evaluation of these three factors, then the level of confidence can be interpreted by policy makers 'as the committee's assessment of the degree to which the answers to the above questions foretell

a resiliency in their credences; as an assessment of how likely their consensus regarding appropriate credences is going to remain fixed in the light of future developments'.

However, it is extremely unclear on what basis Winsberg can argue that under his proposal this is a reasonable interpretation of confidence. Is Winsberg here perhaps inspired by the Bayesians' efforts (discussed in Chapter 6) to try to show that as 'the weight of evidence' increases an agent's credence in a hypothesis tends to become more and more resilient (despite the fact that confidence under Winsberg's proposal cannot be straightforwardly understood as anything close to what Keynes and Bayesians have in mind with the notion of 'the weight of evidence')? Perhaps. However, recall that, as I argued in Section 6.3, this is a reasonable idea only in cases where the agent's credence in a hypothesis is mediated by objective chances and hence their credence in that hypothesis can be interpreted as their expectation of the hypothesis's chance, and it is that expectation that tends to become more and more resilient as 'the weight of evidence' increases. But in this case there is no sense in which the experts' credences in the hypothesis that the ECS is in the range 1.5 °C to 4.5°C are mediated by objective chances.

Recall that under Winsberg's proposal the likelihood metric is supposed to communicate the range of credences that the experts agree it is reasonable to hold in a hypothesis in light of the available evidence. So if the IPCC experts claim that ECS is *likely* to lie in the range 1.5°C to 4.5°C, what this means under Winsberg's proposal is that according to the experts the range of credences that one ought to hold in the hypothesis that the ECS lies in the range [1.5°C , 4.5°C], in light of the available evidence, is (0.66, 1). The range (0.66, 1) is *not* the experts' estimate of the expected chance of the hypothesis. Hence, there doesn't seem to be any reason for supposing that a high level of confidence (somehow based on the evaluation of those three factors) is supposed to give any indication of whether or not the range of credences that the experts assign to that hypothesis will remain fixed in light of future evidence. For instance, Sherwood et al. (2020) seem to be particularly optimistic that further developments in modeling or data acquisition, such as 'improved observation and proxy characterization of

other warm periods in the geological past, which are not yet sufficiently understood' (Sherwood et al. 2020, 107) or continued 'progress in the understanding of cloud feedback mechanisms' (ibid., 106), might in the not too distant future help substantially further down the range of what climate experts consider plausible values for ECS. And if this is right, this should evidently affect the range of credences that the IPCC will assign to the range [1.5°C , 4.5°C] for ECS in light of those developments. In other words, why should the experts, regardless of the evaluation of those three factors today, not substantially change the range of credences that they will assign to the hypothesis that ECS is in the range [1.5°C , 4.5°C] in light of e.g. improved observation and proxy characterization of other warm periods in the geological past? Suppose, for instance, that ECS really is in the range [1.5°C, 4.5°C]. Why should we think that in light of future developments the experts won't consider this range to be "very likely" or "extremely likely"? Or suppose instead that ECS is in fact higher than 4.5°C. Why should we think that, in light of future developments, the experts would still consider this range to be "likely"? If anything, one would hope in this case that in light of future developments the IPCC experts would substantially decrease the credences that they assign to the hypothesis that ECS is in the range [1.5°C, 4.5°C]. In either case, it is very hard to see why an evaluation of confidence which, under Winsberg's proposal, has nothing to do with an evaluation of what evidence the IPCC expect to gather in the future, should give us any indication as to what credences the IPCC experts will assign to a hypothesis in light of future evidence (nor what credences the IPCC experts themselves believe they will assign to a hypothesis in the future).

To be clear, this is not to say that experts may never have good reasons to believe that the credences that they currently assign to a hypothesis is going to remain fixed in light of future evidence. For instance, Consider the following remark in Sherwood et al.'s (2020) recent Bayesian assessment of the evidence relevant to the estimation of the ECS:

> Some of the effects quantified in this paper with the help of GCMs were looked at only with pre-CMIP6 models, and interpretations of

> evidence might therefore shift in the future upon further analysis of newer models, but we would not expect such shifts to be noteworthy unless they involved significant improvements in model skill against relevant observations. (Sherwood et al. 2020, 106)

The above remark suggests that, in this case, the experts do not think that their probabilistic assessment will be greatly affected by the forthcoming evidence from the results of the newer CMIP6 models. But notice that this is mainly due to the nature of the evidence that they *expect* to obtain in the near future, rather than any evaluation of the evidence that they already have.

Hence, to repeat, what I am arguing is that if confidence is supposed to be a summary of those three factors discussed above, as proposed by Winsberg, it really does not look as if the level of confidence has anything to do with the experts' assessment 'of how likely their consensus regarding appropriate credences is going to remain fixed in light of future developments', contrary to what Winsberg suggests.

One way to revise Winsberg's interpretation of confidence as to make it somewhat more plausible might be to argue that the higher the confidence, the more likely the expert's consensus regarding appropriate credences is going to remain *within* the range of credences that they currently assign to the hypothesis. For instance, under this interpretation of confidence, if the experts were to report that a hypothesis is e.g. "likely" with "very high confidence", this would mean that they strongly believe that in light of future developments they will report that that hypothesis is "likely", or "very likely" or "extremely likely". Whereas if the experts were to report that a hypothesis is e.g. "unlikely" with "very high confidence", this would mean that they strongly believe that, in light of future developments, they will report that that hypothesis is "unlikely", or "very unlikely" or "extremely unlikely". However, this interpretation of confidence, although prima facie more plausible, is not at all obvious. This interpretation of confidence is only feasible if the higher the confidence, the more reasons the

experts would have for supposing that future evidence will be in favour of a hypothesis that they currently consider "likely" (or more than likely).[7] But given that the level of confidence, according to Winsberg's proposal, has nothing to do with the nature of the evidence that the experts expect to see in the future, it is really not clear what would justify this view.[8] Hence, overall, I see no reason for suggesting, as Winsberg does, that confidence under his proposal has anything to do with the resiliency of the experts' credences in a hypothesis.

Why confidence and likelihood levels should not interact

Winsberg's proposal gives rise to an another important question to do with the interaction of confidence and likelihood levels: can confidence and likelihood levels interact under this proposal? I will argue that, despite what intuitions one may have about this, confidence and likelihood levels should not be able to interact under Winsberg's proposal. Recall that under this proposal, the likelihood level is determined by the range of credences that almost all, if not all, the relevant experts accept to assign to a hypothesis in light of the available evidence. For instance, if the experts claim that the ESC is "likely" to lie in the range [1.5°C , 4.5°C] what this should mean, under Winsberg's proposal, is that in light of the available evidence, according to almost all, if not all, experts, it is reasonable to assign probability values in the interval [0.66, 1] to the hypothesis that ECS lies in the range [1.5°C , 4.5°C]. Confidence, on the other hand, is supposed to be some sort of subsequent overall evaluation of the available evidence based on those three factors discussed above. So the acceptance of the probability values that should be assigned to a hypothesis comes prior to assigning confidence. Hence, under Winsberg's proposal, it does not look as if one should be able to increase confidence levels by fiddling with the likelihood levels. That is, there is *one* likelihood level, which is determined by the range of satisfactory probability

[7] And also the higher the confidence, the more reasons the experts would have for supposing that future evidence will *not* be in favour of a hypothesis that they currently consider e.g. "unlikely" (or less than unlikely).

[8] In Section 6.3, I argued that it is implausible to argue that as the weight of evidence increases, the resilience of a credence in a range should increase. However, confidence, under Winsberg's proposal, is clearly a different notion from the weight of evidence, since it is affected by factors *other* than the amount of evidence (such as the extent to which the different sources of evidence agree). So given that my scepticism about this revised interpretation of confidence is mainly driven by the discussion of the weight of evidence in Section 6.3, I am aware that I have not given sufficiently strong reasons for rejecting this revised interpretation of confidence.

that should be assigned to a hypothesis agreed on by the experts, and there is *one* confidence level, which is based on some sort of evaluation of the available evidence based on a summary of those three factors. And that should be the end of the story.

But of course, one may wonder: why couldn't the experts increase confidence by simply choosing to report a wider probability interval instead? Surely that should be a way to increase confidence! Let's think about this. Of course, if there were to exist an actual correct probability value that should be assigned to a hypothesis and the IPCC authors were simply unsure about what that value is, then the idea that the IPCC can increase confidence by reporting a wider probability interval would be rather intuitive. A wider probability interval is more likely to contain the correct probability value. Hence, the wider the probability interval, the more confidence the experts should have that the correct probability lies in it. However, under Winsberg's proposal, this is not what is going on here at all. Hence we shouldn't think that the intuition in this example should carry over under Winsberg's proposal. Under Winsberg's proposal, the experts are supposed to agree on a range of credences that should be assigned to a hypothesis in light of the available evidence. Only those credences that are deemed reasonable by the majority of the experts should be included in the interval. For instance, if the experts claim that the ESC is "likely" to lie in the interval [1.5°C , 4.5°C] this must mean, under Winsberg's proposal, that it is *not* the case that the majority of the experts think that it is reasonable to assign any credence in the interval (0, 0.66] to the hypothesis that the ESC lies in the interval [1.5°C , 4.5°C]. But then, it seems to me that reporting a wider probability interval than the one that was agreed on by the experts in light of the available evidence cannot be a viable/coherent way to increase confidence, under Winsberg's proposal, since the experts should not be reporting credences that are not considered to be reasonable by the majority of the experts in light the available evidence.

It just so happens that my views about the lack of interaction between confidence and likelihood levels, under Winsberg's proposal, is in stark contrast with what Winsberg himself might actually think about all this. Winsberg argues that there are many reasons to think that the IPCC conclusions are value laden. One

reason he mentions has to do with the possible interaction between confidence and likelihood levels:

> The IPCC scientists reported "high confidence" in the conclusion that warming greater than 2°C was "likely" under RCP6.0. The "likely" range corresponds to an interval probability assignment of (0.66, 1.0). But the scientists could also have reported a wider probability interval, such as (0.5, 1.0), corresponding to "more likely than not," with even higher confidence, e.g. "very high confidence." Or perhaps they could have reported a narrower interval, such as (0.9, 1.0), corresponding to "very likely," but with less confidence. The question is: what determines which of these representations of uncertainty is communicated? (We assume here that the representations are all consistent with one another.) Without speculating regarding the IPCC example, it seems clear that at least sometimes it is a consideration of the likely applications of an uncertainty report that guide the choice between a wider and more confident report and a narrower and somewhat less confident report. Perhaps a narrower, even if somewhat less confident interval is thought to be more useful for policy makers. In such cases, social values are once again playing a role. (Winsberg 2018, 149)

In the above passage, Winsberg seems to suggest that the idea that likelihood levels can interact with confidence levels makes sense and that the choice of reporting a particular probability interval at a particular confidence level is a choice that that the IPCC experts have to make and one that involves social values. It is possible, however, that Winsberg here is merely referring to the *current* practice of the IPCC authors and hence his comments in the above passage may have nothing to do with whether he thinks that the interaction between confidence levels and likelihood levels is conceptually coherent under his own proposal. But in any case, since Winsberg (2018) does not discuss the interaction of confidence and likelihood under his own proposal, it is important to stress

that under Winsberg's proposal, likelihood levels should not be able to interact with confidence levels.[9] This does *not* mean that social values cannot play a role in what likelihood level to assign. This is because to argue that, under Winsberg's proposal, confidence and likelihood should not interact and hence that the experts should only be able to report one probability interval (rather than any probability interval they feel like, by upgrading/downgrading confidence) is not equivalent to arguing that the acceptance of probabilities in light of the available evidence is not subject to an argument from inductive risk, as Steel (2015) argues. This may very well be. But as discussed earlier, from an inductive risk perspective, accepting probabilities is no less problematic than not accepting them and hence accepting a wider probability interval than the one on which most of the experts agree in light of the available evidence is not a valid way to temper this worry. That is, even if we grant that probabilistic reasoning requires a decision about which probability model, or a set of them, to accept (which will in turn affect which posterior probabilities one accept to assign to a hypothesis in light of the available evidence) and that that decision will necessarily involve some sort of value judgment and hence those values judgments will be 'embedded in whatever posterior probability distribution results from the analysis', this does not support the idea that an IPCC expert panel is free to report whatever probability interval they feel like (by fiddling with confidence levels). The only probability interval that they can and should report, under Winsberg's proposal, is the range of probabilities that are accepted by all or most of the experts and only those.

Before concluding, I should mention that Winsberg's proposal (stripped of misinterpretations of the confidence metric) seems to be close to what Aven (2018) has in mind in the following passage:

[9] Fun fact. In the above passage, Winsberg suggests that the wider the probability interval, the higher the confidence. Under his interpretation of confidence (which as argued in this section can't be right), this would mean that the wider the probability interval, the more strongly the IPCC believe that it will remain fixed in light of future developments. So under this interpretation, the claim that 'we have "high confidence" in the probability interval (0.66, 1) and even higher confidence in the interval (0.5, 1)' would have to be interpreted as something like 'it is rational to assign credences in the range (0.66, 1) and we strongly believe that it will still be rational to assign credences in the range (0.66, 1) in light of future developments & it is rational to assign credences in the range (0.5, 1) and we believe even more strongly that it will still be rational to assign credences in the range (0.5, 1) in light of future developments'. This doesn't strike me as a conceptually coherent claim that the IPCC experts should be able to make.

> To be used in relation to climate change issues, a probability has to be viewed as a subjective (also referred to as a judgemental or knowledge-based) probability, which is conditional on some knowledge. This knowledge can be more or less strong and even erroneous. This fact creates two additional dimensions of risk: firstly, a need for characterising the strength of this knowledge and, secondly, a need for considering surprises relative to the knowledge available. The IPCC works are not explicit on these dimensions, although the former is discussed in relation to statements when referring to evidence and agreement among experts. The problem is, however, [...]: there is no link between the probability judgements and the strength of knowledge judgements in the IPCC framework. From this perspective, the risk analysis science clearly shows that there is such a link and it is essential for understanding risk. (Aven 2018, 292)

Aven, like Winsberg, argues that if the likelihood metric is used to communicate the range of subjective probabilities that the IPCC authors assign to a hypothesis conditional on some knowledge (i.e. conditional on the available evidence), the IPCC should also characterize 'the strength of this knowledge' in some way or another. Unfortunately, Aven does not go into any detail about how such knowledge should be characterized, hence it is hard to tell what he really has in mind and whether he'd be happy with characterizing 'the strength of this knowledge' through something like the confidence metric as in Winsberg's proposal. Nonetheless, Aven is making an important point in the above passage that it is important to stress. However one chooses to characterize and communicate 'the strength of knowledge' on which a probabilistic assignment is conditional, such communication should always come in conjunction with the range of subjective probabilities that the IPCC authors assign to a hypothesis in light of that 'knowledge'. In other words, if the confidence metric is used to communicate 'the strength of knowledge' on which a probabilistic assignment to a hypothesis is conditional on, then one should never assign confidence without also assigning likelihood.

7.3 Mach et al.'s proposal: So long confidence?

Most of the authors of Mach et al.'s (2017) proposal for a new IPCC uncertainty framework, in contrast to Winsberg's proposal (discussed in Section 7.2) and Bradley et al.'s proposal (which I will discuss in Section 7.5), are scientists who have been directly involved with the IPCC assessment and reporting of uncertainties for several years.[10] In light of their experience with the AR5 uncertainty framework, and in response to what they consider a lack of rigour and transparency in the IPCC authors' current usage of the confidence and the likelihood metrics, they propose a new uncertainty framework that is 'is intended to be simultaneously more rigorous and accessible – more straightforward to apply and for readers to understand' (ibid., 10). Despite their intention, however, I will argue that this proposed framework faces most, if not all, the conceptual problems that the current AR5 uncertainty framework faces (discussed extensively in Part 1). Hence, although Mach et al.'s recognition of the some of conceptual problems and ambiguities in the current AR5 uncertainty framework and their ambition to produce a more rigorous and accessible IPCC uncertainty framework is certainly welcome, I will argue that that their proposal does not live up to their ambition.

Mach et al.'s suggested framework includes five terms for describing scientific understanding ('limited', 'emergent', 'medium', 'divergent', 'robust') based on evidence and agreement, as shown below.

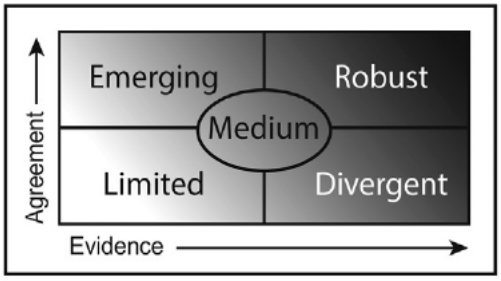

FIGURE 7.1: 'Characterization of scientific understanding' (Mach et al. 2017, 10)

[10]Katherine Mach co-directed the scientific activities of WG II from 2010 till 2015, and is currently a lead author for the AR6. Michael Mastrandrea contributed to the AR4 and was part of the leadership team for the AR5. Patrick Freeman was a coordinating lead author for the AR4 and currently serves as a co-chair of WG II.

It also includes a likelihood scale as in the current uncertainty guide, as shown below.

Likelihood	Probability
Virtually certain	99-100%
Extremely likely	95-100%
Very likely	90-100%
Likely	66-100%
More likely than not	>50-100%
About as likely as not	33-66%
Unlikely	0-33%
Very unlikely	0-10%
Extremely unlikely	0-5%
Exceptionally unlikely	0-1%

FIGURE 7.2: The likelihood scale

However, similarly to Winsberg's proposal discussed in the previous section, the likelihood scale, under this proposal, is supposed to be explicitly based on subjective probabilistic assessments, 'reflecting all plausible uncertainty sources' and should be 'informed by all available evidence, whether it is quantitative, probabilistic, or more diverse' (Mach et al. 2017, 10).

According to Mach et al.'s proposal the above qualitative scientific-understanding terms can either be used as (optional) supplements to likelihood assignments or as a fall back 'when probability cannot be evaluated':

> Where possible and appropriate, experts would assign likelihood or more precise presentations of probability. Scientific-understanding terms would be a supplement or, when probability cannot be evaluated, a fallback. Likelihood could be prioritized especially for key assessment findings, perhaps with more abundant use of scientific-understanding terms in underlying traceable and transparent accounts of evidence and expert judgments. (ibid., 10)

In other words, under this proposal the authors have the option to either assign only likelihood to a finding, or to report likelihood and supplement it with a scientific understanding term, or to only assign a scientific understanding term 'to characterize lower certainty conclusions or broad qualitative conclusions if

the available evidence does not support subjective probabilities' (ibid., 10). Below are some examples that Mach et al. give to show how this would work in practice.

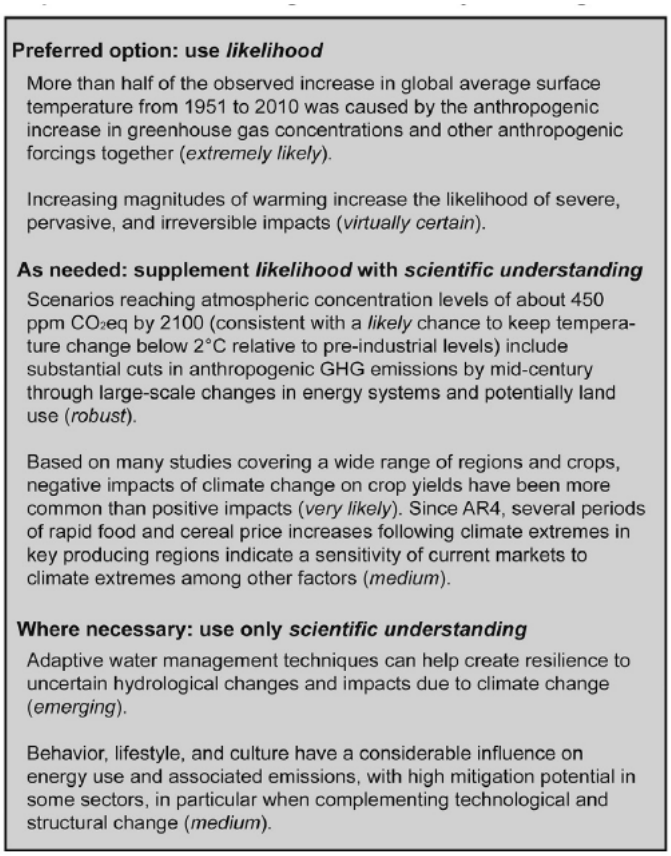

FIGURE 7.3: 'Communication of degree of certainty in findings' (Mach et al. 2017, 10)

What to think of this proposal? One apparent difference between this proposal and the current uncertainty framework for the AR5 is that the confidence metric has disappeared. But is this a significant difference as Mach et al. seem to suggest? In particular, is the relationship between those new qualitative terms for scientific understanding and the likelihood metric clearer than the relationship between "confidence" and "likelihood" in the AR5 uncertainty framework? It seems to me the answer is: no. For Mach et al. have merely replaced the confidence metric with those scientific-understanding terms, without changing any of the aspects that gives rise to the problematic relationship between confidence and likelihood identified in Part 1.

Indeed, notice that under this proposal the authors are once again encouraged to only assign scientific-understanding terms to a finding when 'probability cannot be evaluated'. But then if a scientific-understanding term can be assigned to a finding, without also assigning likelihood, this must mean that those terms cannot be understood as an evaluation of a dimension of uncertainty that *in addition* to the range of credences that the IPCC authors assign to a finding, in light of the available evidence, can be of interest to policy makers (as in Winsberg's proposal discussed above). But if this is *not* their role, then *what is it*? One could perhaps interpret those terms as mere proxies for likelihood assignments, but then there is the question as to why would we need them in the first place since one scale would seem to be enough for the job (i.e. the job of communicating the range of credences that the IPCC authors assign to a hypothesis in light of the available evidence). And if those scientific-understanding terms are *not* supposed to be proxies for likelihood assignments (as arguably is the case) then there is the the question as to what dimension of uncertainty they are supposed to be an evaluation of, and in particular how policy makers should interpret a finding without a likelihood assignment. That is, if a finding to which only a scientific understanding term is assigned is not supposed to tell us anything about what range of subjective probabilities the IPCC assign to a finding, in light of the available evidence, then what is it telling us? [11]

In light of this, it seems to me that Mach et al.'s remark that that this proposal is 'simultaneously more rigorous and accessible – more straightforward to apply and for readers to understand', is not quite justified. For aside from claiming that likelihood is supposed to reflect 'all plausible uncertainty sources' and should be 'informed by all available evidence, whether it is quantitative, probabilistic, or more diverse' (and hence making some *sort* of attempt to clarify the interpretation of the likelihood metric), this proposal suffers from nearly all (if not all) the conceptual problems and ambiguities that the current uncertainty

[11] Recall that another problem with the current AR5 uncertainty framework that I discussed in Part 1 was the puzzling bifurcation of 'evidence' and 'agreement' in the characterization of confidence. Notice that under Mach et al.'s proposal the very same bifurcation of evidence and agreement appears in the characterization of scientific understanding. However, given that Mach et al. do not clarify what the evaluation of evidence and agreement actually depends on, there is not much I can say about it aside from the fact that clearly more would have to be said about what their evaluation depends on under this proposal.

framework suffers from (discussed extensively in Part 1). Hence, for anyone who thinks that those conceptual problems should be resolved, this proposal is not a satisfactory one.

7.4 Some thoughts towards an adequate uncertainty framework: Avoiding the same old mistakes

In this section, I will sketch what a proposal for the IPCC uncertainty framework that satisfies the two desiderata below could look like, while trying not to undermine the honest reporting of the deep uncertainty afflicting studies of climate change and the various challenges that the IPCC authors face in their assessment of it:

1. the framework's fundamental concepts should be clearly defined so that they can be used appropriately and consistently by the IPCC authors in the communication of uncertainty;

2. the use of the framework's fundamental concepts should help the IPCC authors produce findings that are interpretable, relevant and useful for the target audience/s

The proposal I am going to sketch is somewhat similar in spirit to Winsberg's proposal discussed in Section 7.2, but it does nonetheless differ from it in some important ways.

Let's start with 'probability'. As discussed in Chapter 2, the history of the IPCC uncertainty framework is an interesting one: new scales have appeared out of the blue, concepts have changed meaning implicitly if not explicitly. And yet the idea that probability would have to play a role in the communication of uncertainty has never been questioned by the IPCC (every IPCC uncertainty framework that has been produced has always included at least one scale defined in terms of probability ranges). Indeed, probability can undeniably be a very useful concept for the communication of uncertainty. However, it can also be a meaningless, and hence not a very useful, one when used without a clear and understandable interpretation of it.

As discussed in Chapter 2, the very first uncertainty guide (for the AR3) did offer an interpretation of probability. It stated that:

> the probability of an event is the degree of belief that exists among lead authors and reviewers that the event will occur, given the observations, modeling results, and theory currently available. (Moss and Schneider 2000, 36)

This was the first and last time that probability was explicitly defined in an IPCC uncertainty framework and I think it's no coincidence. From the very first uncertainty guide the IPCC authors were not discouraged from using 'frequentist' methods to produce probabilities. Indeed, despite providing the definition of probability above, the first guide itself left room for the authors to decide for themselves when to adopt a 'frequentist' approach' instead, as long this choice was made explicit:

> *authors should explicitly state what sort of approach they are using in a particular case*: if frequentist statistics are used the authors should explicitly note that, and likewise if the probabilities assigned are subjective, that too should be explicitly indicated. Transparency is the key in all cases. (Moss and Schneider 2000, 36; emphasis in the original)

In Chapter 2, we have seen what happened afterwards. In the second revised IPCC uncertainty framework (for the AR4), two quantitative scales appeared instead of one (both were being defined probabilistically). Clearly, transparency was not thought to be enough.

Indeed, transparency is *not* enough! As argued in Section 2.5, some very prominent 'frequentist' (or rather, more accurately, mechanical) methods that are used by the IPCC to assign a probability (or a probability range) to a hypothesis in light of multi-model ensemble results are not conceptually coherent methods for producing objective probabilities (or an estimate of them). Hence, transparency about whether or not one is using some methods rather than others to produce probabilities can't, in my view, be 'the key' if those methods should be not used in the first place.

The reason why I mention all this, is that the methods that the IPCC authors use to assign probabilities to a hypothesis matter. An uncertainty guide which defines probability as subjective and yet does not discourage authors from using conceptually incoherent mechanical methods to produce probabilities is not good enough. Nor of course is an uncertainty guide which *doesn't* define probability at all.

Under my proposal, probability is supposed to be subjective. In particular, as in Winsberg's proposal (and similarly to the first uncertainty guide for the IPCC) the likelihood metric should be used to communicate the range of credences that are satisfactory to almost all, if not all, of the members of a panel i.e. a particular likelihood level assigned to a hypothesis is supposed to communicate the panel's consensus regarding what the range of one's degree of beliefs in a hypothesis ought to be in light of the available evidence. Of course, how the IPCC authors should determine the credences that one ought to have in a hypothesis in light of the available evidence is far from clear in many cases. In particular, as discussed in Chapter 5, evaluating the epistemic import of model consensus in climate science is clearly very hard, hence there is no reason to think that assigning credences to a hypothesis in light of the models' results should be any easier (leaving aside the fact that many scientists might not be Bayesians in the first place). However, if the aim of the IPCC is to truly help policy makers, decision makers and the general public understand what they should think about climate change in light of the available evidence, then the IPCC authors' attempt to evaluate the range of credences that one ought to have in a hypothesis in light of the available evidence is, arguably, better than no attempt at all.

In light of the role of the likelihood metric under this proposal, there are a couple of considerations that I'd like to make. First, since under this proposal a likelihood level is supposed to communicate the range of credences that are satisfactory to the members of a panel and *only those*, it is worth thinking carefully about the ranges of probability that will determine a particular level of likelihood; some ranges might be more pertinent than others. Compare, for instance, the confidence scale in the first uncertainty framework (for the AR3) with the likelihood scale in the current uncertainty framework (for the AR5).

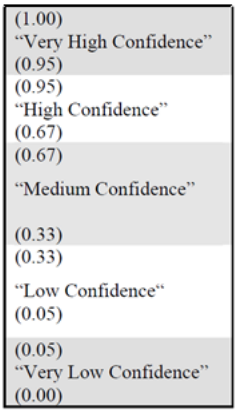

Table 1. Likelihood Scale

Term*	Likelihood of the Outcome
Virtually certain	99-100% probability
Very likely	90-100% probability
Likely	66-100% probability
About as likely as not	33 to 66% probability
Unlikely	0-33% probability
Very unlikely	0-10% probability
Exceptionally unlikely	0-1% probability

FIGURE 7.4: Comparison of the AR3 confidence metric with the AR5 likelihood metric.

Notice that the probability ranges that determine a particular confidence level in the AR3 confidence metric do not overlap with one another, whereas the ones that determine a likelihood level in the AR5 likelihood metric do. I think this difference is significant. And I think that under this proposal there may in fact be good reasons to go back to something like the AR3 confidence metric. As argued in Section 7.2, accepting that it is rational to assign a probability to a hypothesis in light of the available evidence if there are not good reasons for accepting it is no less problematic than *not* accepting that it is rational to assign a probability to a hypothesis in light of the available evidence if there are good reasons for accepting it. So if an IPCC panel does not think it is rational to assign a credence of e.g. 0.99 to a hypothesis in light of the available evidence, that credence should not be included in the ranges of credences that are assigned to a hypothesis. However, if the IPCC authors are given a likelihood metric like the one in the AR5 uncertainty framework, they would be forced to assign a range of credences to a hypothesis that includes e.g. 0.99 even in cases when they only thought it's rational to assign a credence of roughly e.g. 0.7 to a hypothesis. In light of this, I think something like the AR3 confidence metric might be more appropriate for the purpose at hand. Below is my suggestion for what a likelihood metric under this proposal could look like:

Likelihood Scale	
Term	**Credences in a hypothesis**
Virtually certain	0.99-1 credence
Very likely	0.9-0.99 credence
Likely	0.66-0.9 credence
About as likely as not	0.33-0.66 credence
Unlikely	0.1-0.33 credence
Very unlikely	0.01-0.1 credence
Exceptionally unlikely	0-0.01 credence

FIGURE 7.5: The likelihood scale under my proposal.

Second, many important findings in the IPCC are projections of a particular variable in the future, such as this one:

> Increase of global mean surface temperatures for 2081–2100 relative to 1986–2005 is projected to *likely* be in the ranges derived from the concentration-driven CMIP5 model simulations, that is, 0.3°C to 1.7°C (RCP2.6) [. . .] (*very high confidence*) (IPCC 2013b, 20; original emphasis)

As argued in Section 2.5, the most salient feature of this finding is, in my view, the range [0.3°C, 1.7°C]. However, given that the IPCC authors seem to think this range is merely 'likely' I have argued that it is inappropriate to draw so much attention to it. In light of this, under my proposal the IPCC authors should be encouraged to communicate as much as possible the full range of epistemic uncertainty in important projections. For instance, whether the IPCC authors think that under the most positive scenario (RCP2.6) it is 'very unlikely' that by the end of the century increase of global mean surface temperatures will be above 4 °C or whether they think that it is 'very unlikely' that it will be above 2 °C is, in my view, extremely valuable information for policy makers and really everyone else who is concerned about climate change and might want to act on it. Hence, if the IPCC authors think that the range 0.3°C to 1.7°C is 'likely' they should tell us, but this should not preclude them from *also* telling us what range they consider to be e.g 'very likely' in light of the available evidence.

Let's now move on to the confidence metric. I would get rid of it, as it clearly caused too much trouble for its own good. Of course, under my proposal, there might many cases in which the IPCC authors may be forced to assign a likelihood level to a hypothesis (i.e. may forced to tell us their credences in a hypothesis) in light of very limited or perhaps conflicting evidence. But under my proposal this should not preclude them from doing so. For as Moss and Schneider remarked in the very first IPCC uncertainty guide (for the AR3):

> It is certainly true that "science" itself strives for objective empirical information to test theory and models. But at the same time "science for policy" must be recognized as a different enterprise than "science" itself, since science for policy (e.g., Ravetz, 1986) involves being responsive to policymakers' needs for expert judgment at a particular time, given the information currently available, even if those judgments involve a considerable degree of subjectivity. (Moss and Schneider 2000, 36)

Furthermore, the practice of the IPCC authors in their treatment of uncertainties suggests that they often feel more comfortable assigning probabilities to a hypothesis when they can rely on mechanical methods (such as multi-model ensemble methods) to produce them, and less so in other cases. But given that, as argued in Section 2.5, those mechanical methods are not conceptually coherent, there is no justified reason for why this should be the case. Hence, under my proposal the authors should always assign a likelihood level to a hypothesis to express their epistemic uncertainty in that hypothesis, independently of the kind of available evidence. This is a key difference between this and Mach et al.'s (2018) proposal discussed in the previous section.

Having said this, as Winsberg (2018) and Aven (2018) argue, it might be valuable for the authors to have some supplementary qualitative terms for letting policy makers know some aspects of the evidence underpinning their likelihood judgments in a hypothesis, especially when those likelihood judgments are based on very limited or perhaps conflicting evidence. However, in contrast to Winsberg (2018) and Aven (2018), I am less optimistic as to whether those

aspects (whatever we choose them to be) can be neatly summarized into a metric whose evaluation can be of practical value to policy makers and decision makers. As extensively discussed in Chapter 6, the Bayesian has been worrying for a long time about the notion of the weight of evidence underlying a subjective probabilistic assignment to a hypothesis, and as argued in that chapter, the Bayesian (or anyone who uses probability in inductive inference to quantify the degree of belief to assign to a hypothesis given the evidence) has to this day not been able to find an adequate way to measure it, despite acknowledging that it is an important aspect of the evidence. Of course, Keynes's notion/s of the weight of the evidence is/are not quite what Winsberg and Aven seem to have in mind, since according to them the evaluation of the evidence underlying a probabilistic judgment (through the confidence metric) depends on several other factors that are not usually associated with Keynes's notion/s. However, I nonetheless believe that Winsberg's, Aven's and (to a certain extent) also the IPCC's desire to have an additional metric beyond likelihood to characterize 'the strength of knowledge' supporting a probabilistic judgment about a hypothesis is due to essentially the same concern that has driven the Bayesian's perennial attempt to find a solution to the problem of the weight of evidence: the amount and perhaps other just as (if not more) important aspects of the available evidence relevant to a hypothesis are simply not reflected in an agent's credence/s in that hypothesis.

In light of this, I think the best one can hope for is to find some pragmatic but also intelligible way to let policy makers know about some aspects of the evidence underlying a probabilistic judgment that are deemed important. Having qualitative terms such as 'limited', 'emerging', etc., as in Mach et al. proposal (Figure 7.1), could perhaps be an adequate pragmatic solution, but only so long as a little more thought goes into what their evaluation depends on. As argued in Section 2.3, the bifurcation of evidence and agreement in the characterization of confidence in the current AR5 uncertainty framework is conceptually problematic. The fact that the same bifurcation appears in Mach et al.'s characterization of 'scientific understanding' (Figure 7.1) is certainty not reassuring. The problem of how best to characterize the 'strength of knowledge' underpinning

a subjective probability judgment in a hypothesis is clearly a very challenging one, one for which I don't have a solution and that deserves a lot more attention than it has received hitherto. But, however one chooses to characterize it, an important point to keep in mind is that for a proposal to satisfy the second desideratum at the beginning of this section, it should not give the option to assign a qualitative term that is supposed to describe the 'strength of knowledge' underpinning a likelihood assignment without also assigning a likelihood level to that hypothesis.

Below are the key points of the proposal sketched in this section:

1. The likelihood metric (Figure 7.5) should always be used to communicate the range of credences in a hypothesis that are satisfactory to almost all, if not all, of the members of a panel i.e. a particular likelihood level assigned to a hypothesis is supposed to communicate the panel's consensus regarding what the range of one's degree of beliefs in a hypothesis ought to be in light of the available evidence.

2. The IPCC authors should be encouraged to communicate as much as possible the full range of epistemic uncertainty regarding important variables. For instance, the IPCC should be encouraged to tell us both what they consider to be a 'likely' range *and* what they consider to be a 'very likely' range for a variable. This is because the fact that the IPCC authors consider a range for a variable to be 'likely' does not sufficiently help policy makers and decision makers understand what values the IPCC authors believe one is 'warranted' in dismissing for a variable in light of the available evidence.

3. Some additional qualitative terms to describe some important aspects of the evidence underpinning a likelihood assignment might be of some use to policy makers and decision makers to give them some sense of the 'strength of knowledge' underpinning a likelihood assignment. However, what those aspects should be and how they should be evaluated is not at

all obvious. Perhaps merely having a couple of qualitative terms that allows the IPCC authors to highlight when the available evidence is particularly limited and hard to evaluate (for whatever reason e.g. because different lines of evidence support inconsistent hypotheses, or because models do not incorporate important relevant processes etc.) might be enough for the job.

4. However one chooses to address point 3) those additional qualitative terms should *always* be used *in conjunction* with the range of subjective probabilities that the IPCC authors assign to a hypothesis in light of the available evidence, no matter how limited and what not. In other words, those qualitative terms should only be used in conjunction with a likelihood level, never on their own.

Although this is only a sketch of a proposal for a future IPCC uncertainty framework, it is an attempt to give an example of an uncertainty framework that could in principle satisfy the above two desiderata. Having said this, by no means do I think that there could not be other proposals that significantly depart from this one that might also satisfy the above two desiderata. I am also conscious that this proposal leaves open other questions (such as how best to characterize 'the strength of knowledge' underpinning a likelihood assignment). But I nonetheless hope that this sketch can provide some constraints and guidance for future work on developing an adequate IPCC uncertainty framework.

In the next section, I will assess a final proposal by Bradley et al. (2017) that is particularly motivated by desideratum 2. I will argue that it is not an adequate proposal because it fails to satisfy desideratum 1.

7.5 Bradley et al.'s proposal: What decision makers want . . . and how to give it to them without being peer pressured

In a couple of recent papers (Bradley et al. 2017; Helgeson et al., 2018) Bradley, Helgeson and Hill (BHH)[12] offer some suggestions for how to improve and clarify the relationship between confidence and likelihood in the IPCC uncertainty framework. BHH's proposal for how the confidence metric and the likelihood metric should be interpreted and used by the IPCC authors stems mainly from a desire to help clarify what role probability ranges, qualified by confidence judgments, should play in decision making. This is of course an issue of critical importance, one that the IPCC itself should be confronting, given that its main role as an institution is arguably that of informing behaviour and policy. However, the aim of this section is to critically assess the extent to which those suggestions, if taken seriously, would help clarify the interpretation of confidence and likelihood in the IPCC uncertainty framework. Unfortunately, I will argue that their proposal suffers from some serious conceptual problems. Hence, I will conclude that if confidence does have a role to play in the IPCC communication of uncertainty, it can't be the role that they have in mind.

Decision makers are (relatively) comfortable with making decisions when faced with precise probabilities. In this case they can rely on the orthodox normative decision theory, expected utility theory, which prescribes picking the action which maximizes the expected utility relative to the probabilities of the possible states of the world and the utilities.[13] When it comes to imprecise probabilities, decision makers are slightly less comfortable in so far as there is no longer an orthodox normative decision theory on which they can rely to make decisions. However, there are nonetheless a host of possible decision rules that have been offered by decision theorists that can help them in these cases too. A much

[12] For now onward, I will use BHH to refer to the three of them independently of the paper I am quoting from.

[13] Expected utility theory as a normative decision theory is certainly not at all unchallenged, but this is a subject which is beyond the scope of my PhD, so I refer to Steele and Stefánsson (2020) for a comprehensive review of those challenges and possible responses to them.

discussed rule, for instance, is the Maxmin-EU rule, which recommends picking the action with the greatest minimum expected utility relative to the set of probabilities that the decision makers is working with (see, for instance, Gilboa and Schmeidler 1989).[14] Or a less cautious rule is, for instance, the α-Maxmin rule which recommends picking the action with the greatest α-weighted sum of the minimum and maximum expected utilities, again relative to the set probabilities that the decision maker is working with. Under this rule the choice of the relative weight for the minimum and maximum expected utility are supposed to reflect either the decision maker's pessimism or their degree of caution (see, for instance, Ghirardato, Maccheroni, and Marinacci 2004; Binmore 2009).

So far so good. However, some decision theorists have noticed that 'at first pass, the IPCC's uncertainty framework seems far removed from models developed by decision theorists' (Bradley et al. 2017, 503) since what the IPCC delivers are neither precise nor imprecise probabilities. Rather what it delivers is imprecise probabilities *qualified* by qualitative confidence judgments. Hence in light of this, Bradley, Helgeson and Hill worry that it is not sufficiently clear what role those supplementary confidence judgments should play in decision making and that this is a problem for anyone who actually might want to make decisions based on the IPCC findings. Conveniently, however, they do find the one (and only) one decision model that can deal with imprecise probabilities qualified by confidence judgments offered by Hill (2013, 2017). And, in light of this model, they offer some suggestions for how to improve and clarify the relationship of confidence and likelihood in the IPCC uncertainty framework to which I will now turn.

According to BHH's proposal:

> confidence terms attach to the likelihood rather than the outcome directly, two findings can address the same outcome despite using different confidence levels. There is no logical inconsistency in reporting, for example, that the probability of ECS exceeding 6°is 0–.1 (very unlikely) with medium confidence, and 0–.33 (unlikely) with

[14] Although this is a very simple rule to apply, it has been argued that it is overly cautious in so far as the action this rule recommends is not at all affected by the spread of the expected utilities.

> high confidence. The two statements complement one another, together giving an indication of the prevailing trade-off between confidence and precision. Informally, these findings say "We have good evidence that the probability is less than one tenth, and very strong evidence it is no more than one third." On this approach, there is no tension at all between the multiple findings [...]. All of those findings—both original and derived—can be understood as mutually consistent and complementary. (Helgeson et al. 2018, 520)

Notice that BHH's suggestion that it should be possible for the IPCC to assign different likelihood levels qualified by different confidence levels to the same hypothesis is in stark contrast with what I have argued in the previous section. However, since BHH claim 'to provide a simple mathematical model of the confidence-likelihood relationship that resolves outstanding ambiguities while respecting the qualitative nature of the confidence scale' (Helgeson et al. 2018, 518), it is certainly worth hearing out what they have to say.

BHH start with the idea that the assignment of an imprecise probability interval to a hypothesis must be determined by a well defined set of probability distribution functions (pdfs); so for instance 'assigning probability 0–.1 to outcome x means that within the set of pdfs collectively representing authors' uncertainty, the smallest probability given to outcome x by any pdf is 0 and the largest probability given to x by any pdf is .1' (Helgeson 2018, 520). According to BHH's proposal, this set of probability distributions will determine the probability interval that should be assigned to a particular quantity or outcome, but with a caveat. A different level of confidence should be associated 'with its own set of pdfs, where higher confidence sets encompass lower-confidence sets' (Helgeson et al. 2018, 520). So in a nutshell, under this proposal, there are various sets of pdfs, where each set is associated with a level of confidence and determines the probability interval that should be assigned to a hypothesis at that level of confidence. Furthermore, since 'higher confidence sets encompass lower confidence sets' this means that higher confidence sets will always include more pdfs than lower confidence sets; hence, under BHH's proposal, there is a

clear trade-off between confidence and the width of the probability interval that should be assigned to a hypothesis:

> This nesting of sets naturally encodes the trade-off between confidence and precision, since more inclusive sets of pdfs translate to wider probability intervals for any given outcome. Multiple likelihood-plus-confidence findings addressing the same uncertain quantity are mutually consistent if and only if they can be modelled by such a mathematical structure. (Helgeson et al. 2018, 520-521)

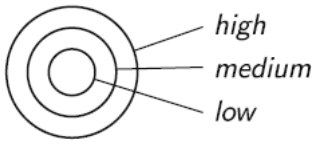

FIGURE 7.6: 'Each confidence level is associated with its own set of probability distributions. The nested structure reflects the trade-off between confidence and the precision of likelihood assignments' (Helgeson et al. 2018, 521)

Bradley et al. (2017, 514-517) give the following toy example to illustrate how this is supposed to work. They ask us to suppose that the author team starts with a well defined set of possible pdfs concerning the value of the equilibrium climate sensitivity (ECS) (where each of these pdfs determines precise probability claims about the values of ECS) and that, 'for concreteness', each pdf is assumed to be lognormally distributed. The author team must then sort those pdfs into what Bradley et al. call a confidence partition, which in this example is assumed to have four elements $\pi = \{M_0, M_1, M_2, M_3\}$. The pdfs in M_0 are supposed to be those considered to be most plausible according to the author team and the pdfs in M_1 'collectively represent a second tier of plausibility. The element M_2 is another step down from there, and M_3 is the bottom of the barrel: all of the pdfs more or less ruled out by the body of research that the experts evaluated' (Bradley et al 2018, 515). This partition of pdfs can then be used to generate a nested family of subsets of pdfs $\{L_0, L_1, L_2, L_3\}$ where L_i is the union

of M_0 through M_i and each L_i is associated with a level of confidence. In this toy example the pdfs have been sorted by the author team in such a way that there are two pdfs in M_0, hence two pdfs in L_0 (since $M_0 = L_0$); and three pdfs in M_1, hence five pdfs in L_1 (since $L_1 = M_0 \cup M_1$) - as is shown in the figure below (the pdfs in M_2 and M_3 are not represented in the figure).

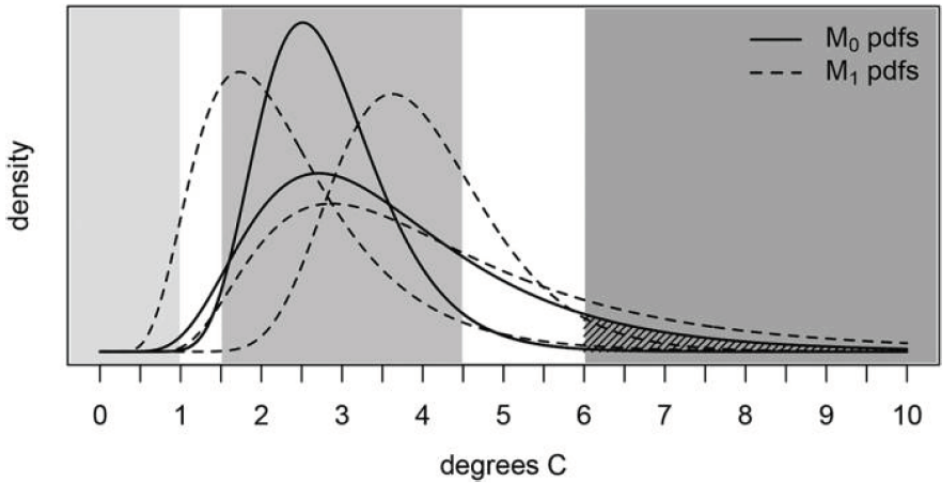

FIGURE 7.7: 'Illustration of a confidence partition consistent with IPCC findings on equilibrium climate sensitivity. The hatched area corresponds to the finding that ECS is very unlikely greater than 6°C (medium confidence)' (Bradley et al. 2017, 516)

Assuming that L_0 corresponds to medium confidence and L_1 to high confidence, the author team is now able to generate various probability statements at those two levels of confidence. For instance, if the experts want to determine the probability interval that should be assigned to the hypothesis that ECS value is greater than 6°at say medium confidence, all they have to do is check what probabilities the various pdfs in L_0 assign to that hypothesis and then report the probability interval bounded by the smallest and largest of those probabilities. In this case L_0 contains two pdfs, where one assigns (nearly) zero probability to the hypothesis that ECS is greater than 6 °and the other assigns just under 0.1 probability. Hence the probability range is roughly $[0 - 0.1]$ which, according to the IPCC terminology, corresponds to 'very unlikely'. Hence the author team can report that "ECS is very unlikely greater than 6 °C (medium confidence)". What if the author team wants to determine the probability interval that should

be assigned to very same hypothesis by with high confidence instead? All they have to do is repeat the same procedure, the only difference being that they now have to take the M_1 pdfs into account in addition to those in M_0 (i.e. they have to consider all pdfs in L_1). In this case the largest probability assigned to the this hypothesis by any of the pdfs in L_1 seems to be a little more than 0.1 but less than 0.33 so in line with IPCC terminology the author team could report that "ESC is unlikely greater than 6°C (high confidence)". The author team can then keep repeating the very same mechanical procedure to determine the probability that should be assigned to all sorts of hypotheses concerning the possible values of ECS at a medium or high confidence. For instance take the hypothesis that the ECS value is within the interval $[1.5 - 4.5]$. This time the smallest probability given by any of the pdfs in L_1 to this hypothesis is a little more than 0.6 where as the highest is nearly one. So in line with the IPCC terminology, the authors could report that "ESC is likely in the interval $[1.5 - 4.5]$ (high confidence)". And so on and so forth.

BHH argue that their proposal clarifies the relationship between confidence and likelihood by providing a principled and transparent method that determines how likelihood and confidence levels interact with one another and that 'systematizes, and enforces consistency among probability intervals assigned to different ranges of a single quantity such as ECS ' (Helgeson et al. 2018, 520) at various different confidence levels. Furthermore, they argue that by showing how it 'can make sense, conceptually to answer the same question at multiple confidence levels' (and at multiple likelihood levels) this proposal can help the IPCC authors give 'a richer picture of scientific knowledge and the added information may be valuable to policy makers and to the public' (Bradley et al. 2017, 519). In particular, they argue that 'by building confidence assessments into a formal belief representation (the nested sets)' (Helgeson et al. 2018, 521), their proposal can help give confidence a clear role in decision making. What they have in mind is a confidence-based decision model proposed by Hill (2013, 2017) which can deal with imprecise probabilities qualified by qualitative (i.e.

ordinal) confidence judgments.[15]

The main considerable challenge BHH acknowledge with their proposal has to do with the calibration of confidence levels between different author teams. That is, for this proposal to work in practice and for it to actually be helpful for decision makers, one would need to make sure that 'what one group means by high confidence is the same as the other' (Bradley et al. 2017, 518) and what one group means by medium confidence is the same as the other (and so on for the other confidence levels). In order to make sure that this is the case, they argue that the IPCC would need to develop 'a proper calibration scale [that] would enable clear and unambiguous formulation and communication of confidence judgments across authors and actors. Were one to take [this] proposal for connecting the IPCC uncertainty language with theories of decision seriously, one major challenge is to develop such a scale' (Bradley et al. 2017, 518).

I agree that were the IPCC to take their proposal seriously, such a calibration scale would have to be developed; however it is extremely unclear what that calibration scale would look like. Moreover, I don't believe the IPCC should take this proposal seriously. My main reason for doubting the adequacy of this proposal owes to its lacking an adequate interpretation of confidence. Intuitively, and in line with the literature on imprecise probabilities, the set of pdfs with which the authors would begin this alleged procedure is supposed to represent all the credence functions that the authors believe are consistent with the available evidence. So on what bases can the authors sort out all those credence functions in a confidence partition in a principled way? The example that Bradley et al. (2017) give suggests that the authors should sort these credence functions based on their confidence in them. But if we understand those credence functions as being all consistent with the available evidence, it is highly unclear why some should be preferred over others and on what bases if so. Without a clear answer to this question it is impossible to understand how this procedure could

[15]The core idea in Hill's decision model is that the probabilistic beliefs an agent adopts will depend on what is at stake in a particular decision problem. That is, what is at stake in a given decision problem will determine the appropriate level of confidence which will in turn determine the set of probability measures taken as the basis for choice. Once the set of probability measures is determined, the agent will once again be in the realm of imprecise probabilities and can choose any of the several decision rules that have been proposed to deal with imprecise probabilities.

even begin. Furthermore, the following remark adds an extra layer of confusion to the problem of how we should interpret confidence under this proposal:

> When used in conjunction with likelihood, we understand confidence to express something like Keynes' (1921/1973) "weight of evidence" behind a probability statement, where the weight he refers to includes the quantity, quality and diversity of evidence underpinning a claim. (Helgeson et al. 2018, 522)

As discussed in Section 6.2, Keynes's notion of 'the weight of evidence' is an ambiguous notion since it can be understood in at least two rather different ways. Given that in this quote they claim that weight includes quantity, quality and diversity of evidence underpinning a probabilistic claim, it is not in fact clear what interpretation of Keynes's weight of evidence they have in mind or whether they are actually referring to Keynes's weight of evidence at all (given that he never explicitly mentions quality and diversity in his various definitions of the weight of evidence). But leaving aside this lack of clarity, and however ambiguous Keynes's notion of the weight of evidence itself is, it is very hard to see why confidence under BHH's proposal has anything to do with Keynes's weight of evidence. The only way, under their proposal, confidence can be understood as expressing something like the 'weight of evidence' , however we choose to understand it, is if we assume that the various sets of pdfs associated with different levels of confidence are based on more or less evidence. And *perhaps* this is what BHH may have in mind after all. As far as the example above is concerned, in order to help the reader understand where the assumption that the pdfs are lognormally distributed comes from, they cite Meehl et al. 2007, sec.10.5.2.1 who write that 'Most studies aiming to constrain climate sensitivity with observations do indeed indicate a similar to lognormal probability distribution of climate sensitivity'. Further down, Meehl et al. provide a summary of the evidence on equilibrium climate sensitivity (Box 10.2), where one can find a host of pdfs concerning the value of ECS obtained from different studies and different lines of evidence. Thus perhaps what BHH have in mind is that the relevant set of pdfs with which the author team should begin this procedure is

the set of all pdfs that have been published in the studies reviewed by the team. Beginning with this set of pdfs, the author team is then supposed somehow to sort them in a confidence partition based on some kind of criterion. But if *this* is what BHH have in mind, it is conceptually flawed. None of the pdfs that are published in the literature can be straightforwardly interpreted as the credence functions of the authors reviewing the available evidence, since each of these pdfs are derived from looking at one particular line of evidence (although not always: sometimes they may be based on the same evidence but derived using different assumptions such as different priors or likelihood functions). But from a Bayesian perspective, the credence functions of the IPCC authors should be based on *all* the evidence available to them, not just on one line of evidence. Of course, combining different lines of evidence using a Bayesian approach is actually very hard, and I am not in anyway suggesting that the authors are able to do that in any sort of rigorous sense.[16] But regardless of how challenging combining different lines of evidence in any rigorous sense can be, the idea that we should simplify that challenge by offering the IPCC authors a mechanical procedure that is conceptually problematic, and that raises all sorts of questions regarding the interpretation of confidence, can't be the right way to deal with this challenge.

Of course, as mentioned earlier, BHH's proposal stems from an urgent need to clarify how the IPCC findings should be interpreted by decision makers (or indeed anyone else!). Moreover, as discussed extensively in Part 1, I also believe it is currently not sufficiently clear how the IPCC findings should be interpreted by any agent who might want to make decisions in light of those findings. However, if the likelihood metric is used by the IPCC authors to communicate the range of credences that, according to the consensus of the author team, one ought to have in a hypothesis in light of all the available evidence (as in my proposal outlined in the previous section), then the likelihood level *will* determine the range of probabilities that an agent should take as input for their decisions. Under my proposal, if the confidence metric has a role to play in the

[16] It is certainly worth noting, however, that there are some recent efforts to do just so at least insofar as equilibrium climate sensitivity is concerned (see, for instance, Stevens et al. 2016 and Sherwood et al. 2020)

IPCC's communication of uncertainty in addition to the likelihood metric, then that role should only be as a *supplementary* evaluation of some of the aspects of the available evidence. Unfortunately, I don't think there is an easy answer to what those aspects should be and to how they should be neatly summarized into a metric, even if a qualitative one (indeed, I further argued we should get rid of the confidence metric and merely have some supplementary terms to describe a few aspects of the available evidence that are deemed important). But this is unfortunately a problem for any scientist that has taken it upon herself to assign credences to a hypothesis in light of the available evidence: the weight of evidence– as Keynes first called it and however we choose to conceptualize it – is, and in my view and always will be, a problem for the Bayesian (or anyone who believes the role of probability in inductive inference is to quantify the degree of belief to assign to a hypothesis given the available evidence). But this problem should not affect the consistency and the coherency of the communication of uncertainty by the IPCC. So if the IPCC thinks that there are important aspects of the evidence that should be communicated in addition to the range of 'acceptable' credences in a hypothesis, then it should be clear what those aspects are. If it is not clear, then it might be better we rid ourselves of confidence (or any supplementary qualitative terms) altogether and avoid unnecessary confusion.

Concluding Remarks

The official aim of the IPCC is 'to provide policymakers with regular scientific assessments on climate change, its implications and potential future risks, as well as to put forward adaptation and mitigation options' and its assessment reports are a key input into the international negotiations to tackle climate change. In light of the potentially catastrophic impacts of climate change on our planet and life as we know it, these assessment reports have (or rather should have) immediate policy implications. There is overwhelming evidence of the disastrous effects of increasing atmospheric levels of greenhouse gases on our planet; in particular, there is overwhelming evidence that temperatures are indeed increasing, will continue to do so and that, in the absence of drastic action to reduce levels of greenhouse gases emissions, severe climate catastrophe will ensue.

However, the climate system is undoubtedly complex, as are its interrelations with humanity. Hence, when it comes to more specific questions beyond these 'big picture' forecasts – How much...? How fast...? Where...? Who...? –, answers are inevitably much harder to furnish. It is because the IPCC does not (and should not) shy away from these questions that it serves as an ideal case study for this thesis's central research question: the question of how to improve our understanding of the challenges involved in the assessment and communication of uncertainty in areas of research deeply afflicted by it, where the assessment and communication of that uncertainty are all the fraughter on account of the studies' immediate policy implications.

We have encountered many challenges in this thesis. As discussed in Part 1, despite the significant effort that has gone into revising and improving the uncertainty framework over the years, IPCC reports continue to suffer from serious conceptual problems, problems that, I argued, have worrying implications

for the IPCC authors' treatment of uncertainties, and the quality of the information provided in the IPCC assessment reports. In Part 2 of this thesis, we have also seen what a formidable challenge the assessment of the epistemic import of model consensus in climate science – and more generally the interpretation of multi-model ensembles' results – really is.

Has this thesis offered any insights that might help us deal with the challenges it identifies? Despite the principally critical nature of this thesis, I believe it has.

For starters, my thesis sheds some light on how to deal with the challenges involved in the assessment of the epistemic import of model consensus in climate science identified in Part 2. There are two key negative conclusions in my thesis that should at the very least steer philosophers on the one hand, and scientists on the other, away from some popular but, I argue, ultimately inauspicious paths for successfully dealing with these challenges.

The first key negative conclusion concerns Winsberg's (2018) recent argument that Schupbach's account of ERA diversity can finally offer us enlightenment on the epistemic import of model robustness in climate science. Winsberg's argument has had an extremely positive reception in the philosophical literature on robustness analysis. According to O'Loughlin (2021, 36), 'Winsberg (2018) convincingly argues that [Schupbach's account] can be applied to climate models.' In reviews of Winsberg's book, Lusk (2019) writes that 'Winsberg's argument is a convincing reconceptualization of robustness analysis in climate science' and Knüsel (2020, 116) that 'Winsberg [...] makes a novel, convincing suggestion for when multiple sources of evidence in favor of a hypothesis are meaningful in climate science.' Despite this extremely positive reception, however, I have argued that Winsberg's argument is flawed and hence cannot shed any light on the epistemic import of model robustness in climate science. As I showed in Section 4.3, Schupbach's (2018) account of ERA diversity seems to fit well and in a straightforward manner with some empirical cases of robustness analysis; however, when one tries to apply Schupbach's account of ERA diversity to model-based robustness analysis, the picture is a good deal more complicated than Schupbach suggests, for its application relies on several non-trivial

assumptions. In Section 5.3, I argued that, whenever the models in an ensemble involve incompatible assumptions about a target system and the hypothesis we are interested in confirming concerns that target system, not all those assumptions can be plausibly satisfied. Hence in all those cases, Schupbach's account is inapplicable. In light of this, I concluded that Winsberg's argument that Schupbach's account of ERA diversity can finally shed light on the epistemic import of model robustness in climate science is flawed. Although this is a negative conclusion, I think it provides a useful lesson for philosophers, especially when it comes to helping scientists evaluate the epistemic import of model robustness. If our aim as philosophers of science is genuinely to help scientists evaluate the epistemic import of model robustness, then we must endeavour to question our intuitions and not allow them to dictate the (often implicit) assumptions that we are willing to accept in order to advance our ultimately unhelpful arguments.

The second key negative conclusion concerns climate scientists' perennial attempt to find a satisfactory measure of independence across climate models, or in other words, a measure of how dissimilar climate models are from one another. In Section 5.4, I argued that what has been driving these efforts is the implicit assumption that the more dissimilar models are from other models in an ensemble, the greater the confidence one should have in the models' consensus. And yet, as I further argued, none of the arguments for the epistemic import of model robustness that I have considered in this thesis can justify why this is a valid assumption. This negative conclusion suggests (despite not conclusively showing) that the frenetic search by climate scientists for a measure of independence able to satisfactorily capture how dissimilar models are from one another not only faces many challenges (some of which I discussed in Section 5.4.1) – it is also misguided. Hence, this also suggests that scientists' current efforts to deal with the extremely challenging problem of the interpretation of climate models' results would be better directed elsewhere. (Where? That is a question that this thesis has, I confess, shed little if any light on.)

Let me now turn to the insights offered by my thesis with respect to the challenges involved in the conceptualization of uncertainty in the IPCC uncertainty framework, discussed in Part 1 of this thesis. In Chapter 1, I argued extensively

that the current IPCC uncertainty framework fails to adequately meet the following two basic desiderata for an adequate uncertainty framework:

1. the framework's fundamental concepts should be clearly defined so that they can be used appropriately and consistently by the IPCC authors in the communication of uncertainty;

2. the use of the framework's fundamental concepts should help the IPCC authors produce findings that are interpretable, relevant and useful for the target audience/s.

In particular, I identified and extensively discussed what I take to be two important conceptual problems in the current IPCC uncertainty framework: the puzzling bifurcation between evidence and agreement in the characterization of 'confidence'; and the lack of an interpretation of the IPCC concepts of 'confidence' and 'likelihood' that is compatible with the IPCC uncertainty guide's recommendations (and thus with the resulting practice of the IPCC authors in their communication of uncertainty). I argued that the ambiguity surrounding the concepts of 'likelihood' and 'confidence' has very serious and worrying implications for both the practice of the IPCC authors in their treatment of uncertainties and the quality of the information provided in the IPCC uncertainty report.

The aim of Part 3 of this thesis was to offer critical reflections on what an adequate IPCC uncertainty framework could in fact look like. In Chapter 7, I evaluated three recent and very different proposals for a new IPCC uncertainty framework that significantly depart from the current one (Winsberg's (2018), Mach et al.'s (2017) and Bradley et al.'s (2017)). After arguing than none of these proposals meet the above two basic desiderata for different reasons, I offered my own tentative sketch of a proposal for a better IPCC uncertainty framework.

I say tentative because it really is: the purpose of my sketched proposal was merely to show what an IPCC uncertainty framework that meets the above two desiderata *could* look like, rather than what the IPCC uncertainty framework *should* look like. Indeed, I hope that Chapter 6's somewhat arcane journey into the troubling notion(s!) of Keynes's weight of evidence – and the light I have

shown it sheds (in the Bayesian's own eyes) on the limitations of an(y) epistemology that envisions the role of probability to be that of quantifying the degree of belief to assign to a hypothesis given the available evidence – has at the very least convinced my reader that what an adequate IPCC uncertainty framework meeting the above two desiderata might look like is anything but obvious. Where do we go from here? That remains an open question.

Bibliography

Abramowitz, Gab et al. (2019). "ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing". In: *Earth System Dynamics* 10.1, pp. 91–105.

Annan, James D and Julia C Hargreaves (2017). "On the meaning of independence in climate science". In: *Earth System Dynamics* 8.1, pp. 211–224.

Aven, Terje and Ortwin Renn (2015). "An evaluation of the treatment of risk and uncertainties in the IPCC reports on climate change". In: *Risk Analysis* 35.4, pp. 701–712.

Baumberger, Christoph, Reto Knutti, and Gertrude Hirsch Hadorn (2017). "Building Confidence in Climate Model Projections: an Analysis of Inferences from Fit". In: *Wiley Interdisciplinary Reviews: Climate Change* 8.3, e454.

Belot, Gordon (2013). "Bayesian orgulity". In: *Philosophy of Science* 80.4, pp. 483–503.

Berger, James O and Robert Wolpert (1988). "The Likelihood Principle and Generalizations". In: *The likelihood principle*. Institute of Mathematical Statistics, pp. 19–64.

Betz, Gregor (2010). "What's the worst case? The methodology of possibilistic prediction". In: *Analyse & Kritik* 32.1, pp. 87–106.

— (2016). "Accounting for possibilities in decision making". In: *the argumentative Turn in policy analysis*. Springer, pp. 135–169.

Binmore, Ken (2008). *Rational decisions*. Princeton University Press.

Bishop, Craig H and Gab Abramowitz (2013). "Climate model dependence and the replicate Earth paradigm". In: *Climate dynamics* 41.3-4, pp. 885–900.

Boe, Julien (2018). "Interdependency in multimodel climate projections: Component replication and result similarity". In: *Geophysical Research Letters* 45.6, pp. 2771–2779.

Bovens, Luc, Stephan Hartmann, et al. (2003). *Bayesian Epistemology*. Oxford University Press.

Bradley, Richard (2009). "Revising incomplete attitudes". In: *Synthese* 171.2, pp. 235–256.

Bradley, Richard, Casey Helgeson, and Brian Hill (2017). "Climate change assessments: confidence, probability, and decision". In: *Philosophy of Science* 84.3, pp. 500–522.

Calcott, Brett (2011). "Wimsatt and the robustness family: Review of Wimsatt's Re-engineering Philosophy for Limited Beings". In: *Biology & Philosophy* 26.2, pp. 281–293.

Cartwright, Nancy (1991). "Replicability, Reproducibility, and Robustness: Comments on Harry Collins". In: *History of Political Economy* 23.1, pp. 143–155.

— (2009). "If no capacities then no credible worlds. But can models reveal capacities?" In: *Erkenntnis* 70.1, pp. 45–58.

Cartwright, Nancy et al. (1994). "Nature's Capacities and their Measurement". In: *OUP Catalogue*.

Chalmers, Alan (2011). "Drawing philosophical lessons from Perrin's experiments on Brownian motion: A response to van Fraassen". In: *The British journal for the philosophy of science* 62.4, pp. 711–732.

Cohen, Jonathan L (1986). "Twelve questions about Keynes's concept of weight". In: *The British journal for the Philosophy of Science* 37.3, pp. 263–278.

Crespo, Ricardo F. (2013). *Theoretical and practical reason in economics: capacities and capabilities*. Springer.

Crupi, Vincenzo and Katya Tentori (2012). "A second look at the logic of explanatory power (with two novel representation theorems)". In: *Philosophy of Science* 79.3, pp. 365–385.

Dittmann, S. et al. (2016). "Hector's Dolphin Movement Patterns in Response to Height and Direction of Ocean Swell". In: *New Zealand Journal of Marine and Freshwater Research* 50.2, pp. 228–39.

Dongen, Noah van, Eric-Jan Wagenmakers, and Jan Sprenger (2020). "A Bayesian Perspective on Severity: Risky Predictions and Specific Hypotheses". In:

Earman, John (1992). *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*.

Earman, John, John T Roberts, and Sheldon Smith (2002). "Ceteris paribus post". In: *Erkenntnis* 57.3, pp. 281–301.

Fairfield, Tasha and Andrew E Charman (2017). "Explicit Bayesian analysis for process tracing: Guidelines, opportunities, and caveats". In: *Political Analysis* 25.3, pp. 363–380.

Feduzi, Alberto (2010). "On Keynes's conception of the weight of evidence". In: *Journal of Economic Behavior & Organization* 76.2, pp. 338–351.

Fitelson, Branden (2001). "A Bayesian Account of Independent Evidence with Applications". In: *Philosophy of Science* 68.S3, S123–S140.

Frigg, Roman (2016). "Chance and determinism". In: *Oxford Handbook of Probability and Philosophy*. Ed. by A. Hájek and C. Hitchcock. Oxford University Press, pp. 460–474.

Frigg, Roman et al. (2014). "Laplace's demon and the adventures of his apprentices". In: *Philosophy of Science* 81.1, pp. 31–59.

Frisch, Mathias (2015). "Predictivism and old evidence: a critical look at climate model tuning". In: *European Journal for Philosophy of Science* 5.2, pp. 171–190.

Funtowicz, Silvio O and Jerome R Ravetz (1990). *Global environmental issues and the emergence of second order science*. Office for Official Publications of the European Communities.

Furrer, Reinhard et al. (2007). "Multivariate Bayesian analysis of atmosphere–ocean general circulation models". In: *Environmental and ecological statistics* 14.3, pp. 249–266.

Gause, George Francis (2019). *The Struggle for Existence: A Classic of Mathematical Biology and Ecology*. Courier Dover Publications.

Ghirardato, Paolo, Fabio Maccheroni, and Massimo Marinacci (2004). "Differentiating ambiguity and ambiguity attitude". In: *Journal of Economic Theory* 118.2, pp. 133–173.

Gilboa, Itzhak, Andrew Postlewaite, and David Schmeidler (2009). *Is it always rational to satisfy Savage's axioms?* Tech. rep.

Gilboa, Itzhak and David Schmeidler (1989). "Maxmin expected utility with non-unique prior". In: *Journal of mathematical economics* 18.2, pp. 141–153.

Glymour, Clark (1980). "Theory and evidence". In:

— (2015). "Probability and the Explanatory Virtues". In: *British Journal for the Philosophy of Science* 66.3, pp. 591–604.

Good, IJ (1985). "Weight of evidence: A brief survey". In: *Bayesian statistics 2* 2, pp. 249–270.

Harris, Margherita (2021). "The epistemic value of independent lies: False analogies and equivocations". In: *Synthese*. URL: https://doi.org/10.1007/s11229-021-03434-8.

Helgeson, Casey, Richard Bradley, and Brian Hill (2018). "Combining probability with qualitative degree-of-certainty metrics in assessment". In: *Climatic Change* 149.3, pp. 517–525.

Hill, Brian (2013). "Confidence and decision". In: *Games and economic behavior* 82, pp. 675–692.

— (2017). "Confidence in beliefs and rational decision making". In: *HEC Paris Research Paper No. ECO/SCD-2018-1258*.

Houkes, Wybo and Krist Vaesen (2012). "Robust! Handle with care". In: *Philosophy of Science* 79.3, pp. 345–364.

Hudson, Robert (2020). "The Reality of Jean Perrin's Atoms and Molecules". In: *The British Journal for the Philosophy of Science* 71.1, pp. 33–58.

Hüttemann, Andreas (2014). "Ceteris paribus laws in physics". In: *Erkenntnis* 79.10, pp. 1715–1728.

IPCC (2000). "Uncertainties in the IPCC TAR: Recommendations to lead authors for more consistent assessment and reporting". In: *Guidance Papers on the Cross Cutting Issues of the Third Assessment Report of the IPCC*. Intergovernmental Panel on Climate Change. Available at http://ipcc.ch, pp. 33–51.

— (2005). *Guidance notes for lead authors of the IPCC fourth assessment report on addressing uncertainties*. Intergovernmental Panel on Climate Change. Available at http://ipcc.ch.

— (2007a). "AR4 Climate change 2007: Impacts, Adaptation, and Vulnerability". In: *Intergovernmental Panel on Climate Change , Cambridge University Press, Cambridge*.

— (2007b). "AR4 Climate Change 2007: The Physical Science Basis". In: *Intergovernmental Panel on Climate Change , Cambridge University Press, Cambridge*.

— (2010a). "Good Practice Guidance Paper on Assessing and Combining Multi Model Climate Projections. In: Meeting Report of the Intergovermental Panel on Climate Change Expert Meeting on Assessing and Combining Multi Model Climate Projections". In: *Working Group I Technical Support Unit, Univeristy of Bern, Bern, Switzerland*.

— (2010b). *Guidance note for lead authors of the IPCC fifth assessment report on consistent treatment of uncertainties*. Intergovernmental Panel on Climate Change. Available at http://ipcc.ch.

— (2013a). *Climate change 2013: The physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

— (2013b). *IPCC factsheet: What is the IPCC?* Intergovernmental Panel on Climate Change. Available at http://ipcc.ch.

— (2021). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

Janzwood, Scott (2020). "Confident, likely, or both? The implementation of the uncertainty language framework in IPCC special reports". In: *Climatic Change* 162, pp. 1655–1675.

Jeffrey, Richard C (1956). "Valuation and acceptance of scientific hypotheses". In: *Philosophy of Science* 23.3, pp. 237–246.

Jeffreys, Harold (1961). *Theory of Probability*. Third. Oxford, England: Oxford.

Jones, Roger N (2011). "The latest iteration of IPCC uncertainty guidance—an author perspective". In: *Climatic Change* 108.4, pp. 733–743.

Joyce, James M (2005). "How probabilities reflect evidence". In: *Philosophical perspectives* 19, pp. 153–178.

Joyce, James M (2010). "A defense of imprecise credences in inference and decision making". In: *Philosophical perspectives* 24, pp. 281–323.

Justus, James (2012). "The Elusive Basis of Inferential Robustness". In: *Philosophy of Science* 79.5, pp. 795–807.

Kandlikar, Milind, James Risbey, and Suraje Dessai (2005). "Representing and communicating deep uncertainty in climate-change assessments". In: *Comptes Rendus Geoscience* 337.4, pp. 443–455.

Kasser, Jeff (2016). "Two Conceptions of Weight of Evidence in Peirce's Illustrations of the Logic of Science". In: *Erkenntnis* 81.3, pp. 629–648.

Katzav, Joel (2014). "The epistemology of climate models and some of its implications for climate science and the philosophy of science". In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 46, pp. 228–238.

Keynes, John Maynard (1921). *A treatise on probability*. Macmillan and Company, limited.

Knüsel, Benedikt (2020). "Philosophy and Climate Science". In: *Ethics, Policy & Environment* 23.1, pp. 114–117.

Knutti, Reto (2018). "Climate model confirmation: From philosophy to predicting climate in the real world". In: *Climate modelling*. Springer, pp. 325–359.

Knutti, Reto, Christoph Baumberger, and Gertrude Hirsch Hadorn (2019). "Uncertainty quantification using multiple models—Prospects and challenges". In: *Computer Simulation Validation*. Springer, pp. 835–855.

Knutti, Reto and Gabriele C Hegerl (2008). "The equilibrium sensitivity of the Earth's temperature to radiation changes". In: *Nature Geoscience* 1.11, pp. 735–743.

Knutti, Reto, Maria AA Rugenstein, and Gabriele C Hegerl (2017). "Beyond equilibrium climate sensitivity". In: *Nature Geoscience* 10.10, pp. 727–736.

Knutti, Reto et al. (2010). "Challenges in combining projections from multiple climate models". In: *Journal of Climate* 23.10, pp. 2739–2758.

Kuorikoski, Jaakko and Aki Lehtinen (2009). "Incredible Worlds, Credible Results". In: *Erkenntnis* 70.1, pp. 119–131.

Kuorikoski, Jaakko, Aki Lehtinen, and Caterina. Marchionni (2010). "Economic Modelling as Robustness Analysis". In: *The British Journal for the Philosophy of Science* 61.3, pp. 541–67.

Kuorikoski, Jaakko, Aki Lehtinen, and Caterina Marchionni (2012). "Robustness Analysis Disclaimer: Please Read the Manual Before Use!" In: *Biology & Philosophy* 27.6, pp. 891–902.

Kuorikoski, Jaakko and Caterina Marchionni (2016). "Evidential Diversity and the Triangulation of Phenomena". In: *Philosophy of Science* 83.2, pp. 227–247.

Lange, Marc (1993). "Natural laws and the problem of provisos". In: *Erkenntnis* 38.2, pp. 233–248.

— (2000). *Natural laws in scientific practice*. Oxford University Press on Demand.

— (2001). "The apparent superiority of prediction to accommodation as a side effect: A reply to Maher". In: *The British journal for the philosophy of science* 52.3, pp. 575–588.

— (2010). "What are Mathematical Coincidences (and Why Does It Matter)?" In: *Mind* 119.474, pp. 307–340.

Leduc, Martin et al. (2016). "Is institutional democracy a good proxy for model independence?" In: *Journal of Climate* 29.23, pp. 8301–8316.

Levins, Richard (1966). "The strategy of Model Building in Population Biology". In: *American Scientist* 54.4, pp. 421–31.

— (1993). "A Response to Orzack and Sober: Formal Analysis and the Fluidity of Science". In: *The Quarterly Review of Biology* 68.4, pp. 547–555.

Lewis, David (1981). "A Subjectivist's Guide to Objective Chance". In: *IFS: Conditionals, Belief, Decision, Chance and Time*. Ed. by William L. Harper, Robert Stalnaker, and Glenn Pearce. Dordrecht: Springer Netherlands, pp. 267–297.

Lipton, Peter (1999). "All else being equal". In: *Philosophy* 74.288, pp. 155–168.

Lisciandra, Chiara (2017). "Robustness Analysis and Tractability in Modeling". In: *European Journal for Philosophy of Science* 7.1, pp. 79–95.

Lloyd, Elisabeth A (2009). "Varieties of support and confirmation of climate models". In: *Proceedings of the Aristotelian Society* 83.1, pp. 213–232.

— (2010). "Confirmation and robustness of climate models". In: *Philosophy of Science* 77.5, pp. 971–984.

Lloyd, Elisabeth A (2015). "Model robustness as a confirmatory virtue: The case of climate science". In: *Studies in History and Philosophy of Science Part A* 49, pp. 58–68.

Lusk, Greg (2019). "Philosophy and Climate Science, by Eric Winsberg". In: *The British Journal for the Philosophy of Science Review of Books*. URL: http://www.thebsps.org/reviewofbooks/lusk-on-winsburg/.

Mach, Katharine J et al. (2017). "Unleashing expert judgment in assessment". In: *Global Environmental Change* 44, pp. 1–14.

Manning, Martin R (2006). "The Treatment of Uncertainties in the Fourth IPCC Assessment Report". In:

Mason, Robert L, Richard F Gunst, and James L Hess (2003). *Statistical design and analysis of experiments: with applications to engineering and science*. Vol. 474. John Wiley & Sons.

Mastrandrea, Michael D and Katharine J Mach (2011). "Treatment of uncertainties in IPCC Assessment Reports: past approaches and considerations for the Fifth Assessment Report". In: *Climatic Change* 108.4, pp. 659–673.

Mastrandrea, Michael. D. et al. (2011). "The IPCC AR5 guidance note on consistent treatment of uncertainties: A common approach across the working groups." In: *Climate Change* 108, pp. 675–691.

Mayo, Deborah G (1986). "Cartwright, causality, and coincidence". In: *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*. Vol. 1986. 1. Philosophy of Science Association, pp. 42–58.

— (2018). *Statistical inference as severe testing*. Cambridge: Cambridge University Press.

McGrew, Timothy (2003). "Confirmation, heuristics, and explanatory reasoning". In: *The British Journal for the Philosophy of Science* 54.4, pp. 553–567.

McMullin, Ernan (1985). "Galilean Idealization". In: *Studies in History and Philosophy of Science Part A* 16.3, pp. 247–73.

Morey, Richard D et al. (2016). "The fallacy of placing confidence in confidence intervals". In: *Psychonomic bulletin & review* 23.1, pp. 103–123.

Nielsen, Michael (2020). "Deterministic convergence and strong regularity". In: *The British Journal for the Philosophy of Science* 71.4, pp. 1461–1491.

Odenbaugh, Jay and Anna Alexandrova (2011). "Buyer Beware: Robustness Analyses in Economics and Biology". In: *Biology & Philosophy* 26.5, pp. 757–71.

O'Donnell, Rod M (1989). *Keynes: Philosophy, economics and politics: The philosophical foundations of Keynes's thought and their influence on his economics and politics*. Springer.

O'Loughlin, Ryan (2021). "Robustness reasoning in climate model comparisons". In: *Studies in History and Philosophy of Science Part A* 85, pp. 34–43.

Orzack, Steven Hecht and Elliott Sober (1993). "A Critical Assessment of Levins's 'The Strategy of Model Building in Population Biology' (1966)". In: *The Quarterly Review of Biology* 68.4, pp. 533–46.

Parker, Wendy (2014). "Values and uncertainties in climate prediction, revisited". In: *Studies in History and Philosophy of Science Part A* 46, pp. 24–30.

Parker, Wendy S. (2006). "Understanding Pluralism in Climate Modeling". In: *Foundations of Science* 11.4, pp. 349–368.

— (2009). "Confirmation and adequacy-for-purpose in climate modeling". In: *AGUFM* 2009, GC34A–02.

— (2011). "When climate models agree: The significance of robust model predictions". In: *Philosophy of Science* 78.4, pp. 579–600.

— (2013). "Ensemble Modeling, Uncertainty and Robust Predictions". In: *Wiley Interdisciplinary Reviews: Climate Change* 4.3, pp. 213–223.

— (2020). "Model evaluation: An adequacy-for-purpose view". In: *Philosophy of Science* 87.3, pp. 457–477.

Parker, Wendy S and James S Risbey (2015). "False precision, surprise and improved uncertainty assessment". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 373.2055, p. 20140453.

Peirce, Charles S. (1878). "In Writings of Charles S. Peirce: A Chronological Edition, Volume 3: 1872–1878". In: Indiana University Press. Chap. The Probability of Induction.

Perrin, Jean (1916). *Atoms, translated by Hammick D. L.*

Pezzullo, John C. (2005). "Tolerance Intervals for Normal Distribution". In: *Interactive Statistics page*. URL: https://statpages.info/tolintvl.html.

Pirtle, Zachary, Ryan Meyer, and Andrew Hamilton (2010). "What does it mean when climate models agree? A case for assessing independence among general circulation models". In: *environmental science & policy* 13.5, pp. 351–361.

Popper, Karl (1959). *The Logic of Scientific Discovery*. Routledge.

Psillos, Stathis (2011). "Moving molecules above the scientific horizon: On Perrin's case for realism". In: *Journal for General Philosophy of Science* 42.2, pp. 339–363.

Räz, Tim (2017). "The Volterra Principle Generalized". In: *Philosophy of Science* 84.4, pp. 737–760.

Rehg, William and Kent Staley (2017). ""Agreement" in the IPCC Confidence measure". In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 57, pp. 126–134.

Reutlinger, Alexander et al. (2021). "Ceteris Paribus Laws". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2021. Metaphysics Research Lab, Stanford University.

Rice, Collin (2019). "Models don't decompose that way: A holistic view of idealized models". In: *The British Journal for the Philosophy of Science* 70.1, pp. 179–208.

Risbey, James S and Milind Kandlikar (2007). "Expressions of likelihood and confidence in the IPCC uncertainty assessment process". In: *Climatic change* 85.1, pp. 19–31.

Roussos, Joe, Richard Bradley, and Roman Frigg (2020). "Making confident decisions with model ensembles". In: *Philosophy of Science*.

Rudner, Richard (1953). "The scientist qua scientist makes value judgments". In: *Philosophy of science* 20.1, pp. 1–6.

Runde, Jochen (1990). "Keynesian uncertainty and the weight of arguments". In: *Economics & Philosophy* 6.2, pp. 275–292.

Sanderson, Benjamin M, Reto Knutti, and Peter Caldwell (2015). "Addressing interdependency in a multimodel ensemble by interpolation of model properties". In: *Journal of Climate* 28.13, pp. 5150–5170.

Sauve, Alix MC, Rachel A Taylor, and Frédéric Barraquand (2020). "The effect of seasonal strength and abruptness on predator–prey dynamics". In: *Journal of theoretical biology* 491, p. 110175.

Schnee, M. E. and A. J. Ricci (2003). "Biophysical and Pharmacological Characterization of Voltage-Gated Calcium Currents in Turtle Auditory Hair Cells". In: *The Journal of Physiology* 549.3, pp. 697–717.

Schrenk, Markus (2007). "Can Capacities Rescue us from Ceteris Paribus Laws?" In: *Dispositions and Causal Powers*. Aldershot: Ashgate, pp. 221–247.

Schupbach, Jonah N. (2018). "Robustness Analysis as Explanatory Reasoning". In: *The British Journal for the Philosophy of Science* 69.1, pp. 275–300.

Schupbach, Jonah N and Jan Sprenger (2011). "The Logic of Explanatory Power". In: *Philosophy of Science* 78.1, pp. 105–127.

Sherwood, SC et al. (2020). "An assessment of Earth's climate sensitivity using multiple lines of evidence". In: *Reviews of Geophysics* 58.4, e2019RG000678.

Skyrms, Brian (1977). "Resiliency, propensities, and causal necessity". In: *The Journal of Philosophy* 74.11, pp. 704–713.

Spiegelhalter, David J and Hauke Riesch (2011). "Don't know, can't know: embracing deeper uncertainties when analysing risks". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 369.1956, pp. 4730–4750.

Stainforth, David A et al. (2007a). "Confidence, uncertainty and decision-support relevance in climate predictions". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 365.1857, pp. 2145–2161.

Stainforth, David A et al. (2007b). "Issues in the interpretation of climate model ensembles to inform decisions". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 365.1857, pp. 2163–2177.

Steel, Daniel (2015). "Acceptance, values, and probability". In: *Studies in History and Philosophy of Science Part A* 53, pp. 81–88.

Steele, Katie (2012). "The scientist qua policy advisor makes value judgments". In: *Philosophy of Science* 79.5, pp. 893–904.

Steele, Katie and H. Orri Stefánsson (2020). "Decision Theory". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2020. Metaphysics Research Lab, Stanford University.

Stensrud, David J (2009). *Parameterization schemes: keys to understanding numerical weather prediction models*. Cambridge University Press.

Stevens, Bjorn et al. (2016). "Prospects for narrowing bounds on Earth's equilibrium climate sensitivity". In: *Earth's Future* 4.11, pp. 512–522.

Tebaldi, Claudia, Julie M Arblaster, and Reto Knutti (2011). "Mapping model agreement on future climate projections". In: *Geophysical Research Letters* 38.23.

Van Der Bles, Anne Marthe et al. (2019). "Communicating uncertainty about facts, numbers and science". In: *Royal Society open science* 6.5, p. 181870.

Vandermeer, John (1996). "Seasonal isochronic forcing of Lotka Volterra equations". In: *Progress of Theoretical Physics* 96.1, pp. 13–28.

Vezér, Martin A (2016). "Computer models and the evidence of anthropogenic climate change: An epistemology of variety-of-evidence inferences and robustness analysis". In: *Studies in History and Philosophy of Science Part A* 56, pp. 95–102.

Weisberg, Michael (2006). "Robustness Analysis". In: *Philosophy of Science* 73.5, pp. 730–42.

— (2012). *Simulation and similarity: Using Models to Understand the World*. Oxford University Press.

Weisberg, Michael and Kenneth Reisman (2008). "The Robust Volterra Principle". In: *Philosophy of Science* 75.1, pp. 106–131.

Wimsatt, William C (1981). "Robustness, Reliability, and Overdetermination". In: *Scientific Inquiry and the Social Sciences*, pp. 124–163.

Winsberg, Eric (2012). "Values and uncertainties in the predictions of global climate models". In: *Kennedy Institute of Ethics Journal* 22.2, pp. 111–137.

— (2018a). "What Does Robustness Teach Us in Climate Science: A Re-Appraisal". In: *Synthese*, pp. 1–24.

— (2018b). *Philosophy and Climate Science*. Cambridge University Press.

Woodward, Jim (2006). "Some varieties of robustness". In: *Journal of Economic Methodology* 13.2, pp. 219–240.

Wüthrich, Nicolas (2017). "Conceptualizing uncertainty: an assessment of the uncertainty framework of the Intergovernmental Panel on Climate Change". In: *EPSA15 Selected Papers*. Springer, pp. 95–107.

Wynne, Brian (1992). "Uncertainty and environmental learning: reconceiving science and policy in the preventive paradigm". In: *Global environmental change* 2.2, pp. 111–127.

Yeomans, Martin R. et al. (2009). "Effects of Energy Density and Portion Size on Development of Acquired Flavour Liking and Learned Satiety". In: *Appetite* 52.2, pp. 469–78.