The London School of Economics and Political Science

POLICYMAKING UNDER SCIENTIFIC UNCERTAINTY

JOE ROUSSOS

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

May 2020

Joe Roussos: *Policymaking under scientific uncertainty*

DECLARATION

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

I confirm that material from sections 5.3–5.5 is my contribution to a paper co-authored with Roman Frigg and Richard Bradley. Section 5.5 includes material developed by Roman Frigg. The material in this section is 60% my own, and the remainder is included for continuity.

The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent. I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that this thesis consists of 84,412 words.

ABSTRACT

Policymakers who seek to make scientifically informed decisions are constantly confronted by scientific uncertainty and expert disagreement. This thesis asks: how can policymakers rationally respond to expert disagreement and scientific uncertainty? This is a work of non-ideal theory, which applies formal philosophical tools developed by ideal theorists to more realistic cases of policymaking under scientific uncertainty.

I start with Bayesian approaches to expert testimony and the problem of expert disagreement, arguing that two popular approaches—supra-Bayesianism and the standard model of expert deference—are insufficient. I develop a novel model of expert deference and show how it can deal with many of these problems raised for them. I then turn to opinion pooling, a popular method for dealing with disagreement. I show that various theoretical motivations for pooling functions are irrelevant to realistic policymaking cases. This leads to a cautious recommendation of linear pooling. However, I then show that any pooling method relies on value judgements, that are hidden in the selection of the scoring rule.

My focus then narrows to a more specific case of scientific uncertainty: multiple models of the same system. I introduce a particular case study involving hurricane models developed to support insurance decision-making. I recapitulate my analysis of opinion pooling in the context of model ensembles, confirming that my hesitations apply. This motivates a shift of perspective, to viewing the problem as a decision theoretic one. I rework a recently developed ambiguity theory, called the confidence approach, to take input from model ensembles. I show how it facilitates the resolution of the policymaker's problem in a way that avoids the issues encountered in previous chapters.

This concludes my main study of the problem of expert disagreement. In the final chapter, I turn to methodological reflection. I argue that philosophers who employ the mathematical methods of the prior chapters are *modelling*. Employing results from the philosophy of scientific models, I develop the theory of normative modelling. I argue that it has important methodological conclusions for the practice of formal epistemology, ruling out popular moves such as searching for counterexamples.

For Mercédès, 1955–2018

ACKNOWLEDGMENTS

Thank you to my supervisors, Roman Frigg and Richard Bradley, who have been wonderful mentors, offering guidance, support and encouragement as needed.

I can't imagine a better PhD experience than I have had at the LSE. The department has helped me develop intellectually and personally in a way that I will always appreciate. Thanks to the faculty who were always willing to talk about their work and mine, especially Anna Mahtani—chapter 7 developed as a conversation with her about what we're up to when we do formal epistemology. Liam Kofi Bright, Christian List, and Bryan Roberts were also enthusiastic and friendly in their engagements with my work, and for that I am very grateful.

Thanks to all my fellow PhD students. You were great colleagues and friends, and I'm sad to be leaving. Your engagement in the PhD Seminar was extremely helpful in formulating this work. The members of the Formal Epistemology Reading Group were great companions in learning about this field: Charles Beasley, Chloé de Canson, Margherita Harris, and Ko-Hung Kuan. Past LSE graduates were also a big part of my PhD experience. I benefited from discussions about chapter 7 with James Nguyen and Seamus Bradley. Katherine Furman is the reason I came to the LSE, without her friendship and encouragement this may never have happened.

Life at LSE wasn't all about work. Thanks to Charles, Nicolas Côté, Ko-Hung, and James Wills for discussions about anything and everything over lunch. Thanks to James, Nick Makins, and Richard for the much needed exercise and fun provided by our squash group. Friday afternoons at the pub were a great part of life as an LSE PhD student, and I'll say a special thanks to Mike Otsuka for being a stalwart supporter of that key social institution.

Philosophy is a surprisingly social practice, and I have really benefited from that. I presented material from this thesis at various conferences and workshops over the past three years, and I got valuable engagement, criticism and encouragement from many great philosophers. Thanks to all of them.

A special thank you to Tom Philp, who was a crucial guide to the hurricane insurance world and the scientific literature in particular. His friendly and patient explanations of all things hurricane-related were invaluable in working on chapters 4 and 5.

Thanks to my family, and all of my friends, for countless acts of love, support, and companionship. Tim and Anna, you are the best of friends. Long live the WSC.

Most importantly, thanks to Chiara: partner, friend, and proofreader extraodinaire. These have been some of the best years of my life because of you.

CONTENTS

1 INTRODUCTION 13
- 1.1 Seeking consensus 17
- 1.2 Decision-making with unresolved uncertainty 24
- 1.3 Outline 28

2 EXPERT DEFERENCE AND DISAGREEMENT 37
- 2.1 Introduction 37
- 2.2 Supra-Bayesianism 40
- 2.3 Expert deference as a constraint on priors 47
- 2.4 Expert deference as a belief revision schema 53
  - 2.4.1 First pass 57
  - 2.4.2 Belief revision theory 59
  - 2.4.3 Three kinds of deference 63
  - 2.4.4 Summary so far 74
- 2.5 Awareness growth 75
- 2.6 Disagreeing experts 88
  - 2.6.1 Conclusion 90
- 2.7 Addendum: Belief revision by divergence minimisation 92

3 OPINION POOLING 105
- 3.1 Introduction 105
- 3.2 Linear pooling 108
  - 3.2.1 General motivation 109
  - 3.2.2 Axiomatic characterisation 124
- 3.3 Geometric pooling 127
  - 3.3.1 Axiomatic characterisation 128
- 3.4 Weights 135
  - 3.4.1 Choosing a scoring rule 139
  - 3.4.2 The role of values in choosing a rule 145

4 SCIENTIFIC MODEL ENSEMBLES 155
- 4.1 Introduction 155
- 4.2 Ensembles of simulation models 156
- 4.3 Decision-support with model ensembles 162
- 4.4 Why average? 166
- 4.5 Case study: hurricane insurance in the North Atlantic 169
  - 4.5.1 History of the RMS MTR 177
- 4.6 Problems with averaging in the hurricane case 178

5 DECIDING WITH MODEL ENSEMBLES 191
5.1 Introduction . . . 191
5.2 Decision Theory primer . . . 193
5.2.1 Standard decision theory . . . 194
5.2.2 Imprecise probability and decision theory . . . 196
5.3 Confidence theory . . . 204
5.3.1 Deciding with confidence . . . 212
5.3.2 Applying the confidence approach to policy decisions . . . 220
5.4 Application: hurricane insurance . . . 223
5.4.1 Conclusion . . . 234
5.5 Methods for Constructing Nested Sets . . . 238
5.5.1 Cases with an Identified Centre . . . 239
5.5.2 Cases without an Identified Centre . . . 242
5.6 Objections and concerns . . . 247

6 CONCLUSION 253

7 CODA: FORMAL PHILOSOPHY AS MODELLING 263
7.1 Introduction . . . 263
7.2 The target . . . 266
7.3 Scientific models . . . 272
7.4 The methodology of modelling . . . 278
7.5 Normative models . . . 280
7.5.1 The architectural model . . . 282
7.5.2 Idealisation . . . 285
7.5.3 Representation . . . 288
7.5.4 Purpose, scope and criticism . . . 291
7.6 Aside: what else could we be doing? . . . 298
7.7 Methodology and inference . . . 300
7.7.1 Securing normative inferences . . . 303
7.8 Conclusion . . . 308

BIBLIOGRAPHY 311

1

INTRODUCTION

How should we make policy decisions in the face of scientific uncertainty? In this thesis I seek to address some philosophical aspects of this question. This is a work of philosophy, and the issues I deal with are inevitably both abstract and quite specific. I will therefore start with a "big picture" introduction to the issue of policy decision-making in the presence of scientific uncertainty. My aim is to show where this theoretical work is intended to have downstream impact, and to motivate for approaching the topic as I have.

Policymakers are increasingly reliant on input from experts in making their decisions. This is a good thing, reflecting a desire to make policy that is informed by the best evidence available. Often, policy engagement with expertise looks something like this: a policymaker needs an answer to a question in order to inform decision-making; there is a relevant body of expertise and a community of experts; however, there is no ready-made or widely endorsed answer to this particular question. Rather than attempting to master this expert domain, or asking an individual expert, policymakers engage in various processes for assessing the position of the community of experts. These processes range from panels with a handful of experts to large collaborative assessments.

Unfortunately, experts are often uncertain, and they regularly disagree. Uncertainty is a problem for decision-making.[1] Most straightforwardly, it obscures which action is best. If we were certain what would happen, we could respond appropriately; to the extent that we are uncertain, other options begin to look plausible and we risk making the wrong decision. Reducing uncertainty is therefore an important practical goal—in addition to having obvious epistemic benefits. Uncertainty also generates problems for policy decisions indirectly: politicians who are unwilling to act on contentious issues may play up uncertainty in order to motivate for delay (Oppenheimer et al., 2019, p. 12); bad actors who wish to suppress policy regulating their harmful behaviour may manufacture or exaggerate uncertainty in order to obscure which actions are best (Oreskes and Conway, 2010).

Policy decision-makers therefore need tools for "managing" uncertainty. This might mean reducing it, or working with it.

The issues that I discuss in this thesis are of relevance to all policy engagements with uncertainty, but I will focus largely on two sorts of cases. In the first sort, a panel of experts is convened to inform a policymaker. They are asked to provide their opinions on a range of questions, and they disagree. In the second sort, scientists use a collection of scientific models—also called an ensemble—to make predictions and provide answers to policy-relevant questions. These models produce contrary outputs.

[1] In this introduction, I will use "uncertainty" as an umbrella term that includes disagreement between experts, as this is a source of uncertainty for the policymaker.

Models and model ensembles will be introduced later on, but here is a brief introduction to expert panels and expert elicitation.[2] An expert panel is a small group of experts (between, say, 4 and 20), convened to provide answers and advice on a specific question. Panels are convened by policymakers, often with the help of specialist facilitators. To start, the policymakers define a set of problems, identify the relevant area of expertise and a few candidate experts. Often this process is iterative: experts identified early are engaged to help refine the problem-definition, tighten the target area of expertise and identify more relevant experts. A format is agreed for the process of gathering information from the panel—often called the "elicitation." Experts are invited to join the panel, and sessions are held. There are different approaches to facilitating these sessions, a few of which will be mentioned below. Some seek to develop a unified, consensus position; others seek to merely elicit the opinions of the gathered experts and to present them (typically in summary form) to the policymakers. Sessions can last from a few days to a few weeks. Some panels are convened for a single session, others are convened regularly over a number of years.

For ease of discussion, here is a toy case that I will refer back to throughout this thesis.

Case 1. *Ade is a policymaker, trying to decide how to enhance Thames flood defences for the next fifty years.*[3] *He wishes to use the best scientific advice available to determine the likelihood that the Thames will rise more than*

[2] Following descriptions in (Cooke, 1999), (Pulkkinen and Simola, 2000) and (Cooke and Goossens, 2000).

[3] Ade is a Nigerian name, pronounced uh-DAY rather than AID.

50cm—which would require new barriers. He convenes a panel of experts. The 10 experts disagree, offering a wide range of answers, from unlikely to very likely.

Some preliminary philosophical framing will help set up future discussions. I will assume that Ade is a novice in this domain, which is to say that "he is not in a position to evaluate the target experts by using his own opinion; at least, he does not think he is in such a position. The novice either has no opinions in the target domain, or does not have enough confidence in his opinions in this domain to use them in adjudicating or evaluating the disagreement between the rival experts" (Goldman, 2001, p. 90). This is common: policymakers may gain some familiarity with a field such as climate science through years of exposure to the topic, but they are not specialists in the technical details of the science, nor are they typically equipped to adjudicate disputes of theory choice, modelling technique or statistical analysis.

My discussion will be normative, concerned with the *rational* options open to Ade. It will start out in epistemology, focused on what Ade should believe, and then become decision theoretic, focused on the choice-related aspects of his situation.

Some epistemological preliminaries: Throughout, I will assume that expert testimony can—and often does—warrant belief. A truth-seeking layperson will do well by adopting the beliefs that experts profess in their domains (Hardwig, 1985, 1991). In addition to assuming this is epistemically justified, I note that it is pragmatically a matter of necessity due to our limited cognitive resources. Individual agents can expect to achieve expertise in at most a few domains, and for the remainder they will do best to defer to experts rather than attempting to reason for themselves. I recognise that this passes over substantial debates in contemporary epistemology, but one must choose one's battles. I have found plenty of philosophical interest in thinking about *how* one should use expert testimony to inform one's beliefs, rather than *whether* one should.

1.1 SEEKING CONSENSUS

One major approach to managing uncertainty in policy situations is to attempt to remove or resolve uncertainty; to build a *consensus* for the purpose of policymaking. Consensus commonly refers to (near) universal agreement, though it is occasionally also used to refer to a kind of collective acceptance.[4] Expert consensus is often taken to confer special epistemic status on a claim. In policy contexts in particular, it has come to be taken to indicate knowledge that is "decision-ready" (Kennel, 2015). Consensus is taken to indicate that experts agree that this knowledge is *important and settled enough* to form a basis for policy (Oppenheimer et al., 2019, p. 11).

Consensus-seeking and -building is one of the major aims of scientific assessments such as those of the Intergovernmental Panel on Cli-

[4] For example, Miller develops an analysis of consensus on which "consensus amongst group G that *p*" is not the same as "most members of G believe that *p*," but rather means something like "most members of G use *p* as if it were true in their reasoning, and endorse it as the position of G, although they may not personally believe it or hold that it is true" (I am paraphrasing Miller, 2013, pp. 1295-7). It seems plausible that many consensus positions in science are of this sort, and my discussion below holds equally well for Miller's definition.

mate Change (IPCC) and the USA's National Acid Precipitation Assessment Program (NAPAP).[5] A much smaller-scale procedure with the same aim is the Delphi method for expert elicitation. This is a process of iterated information-sharing and discussion, aimed at investigating sources of disagreement and eliminating them.

My focus in this work is on methods which *do not* seek or construct consensus. Before introducing them, I will briefly discuss reasons to doubt that consensus-seeking is the right approach. This is not intended to be exhaustive or definitive, but to illustrate why we might want alternatives to this attractive-sounding option.

In short, consensus is difficult to achieve. Shorter-timeline methods for "building" consensus introduce problematic forcing mechanisms which undermine the value of the resulting "consensus." The slower, more reliable assessment processes can sometimes take too long to be useful for decisions. In either case, a focus on consensus impedes decision-making on urgent issues where action is needed.

Short-term consensus building is unreliable

The Delphi method can be thought of as an attempt to "hot house" the social practice of science, by forcing debate and convergence in a controlled setting. A review of some of its (well-known) problems highlights what is difficult about building consensus on a short timeline.

[5] I don't want to imply that the IPCC in particular exists *only* to seek or present consensus. But the summary for policymakers, in particular, attempts to present only results that are widely agreed upon and what is included in the summary is a matter of political contention, as discussed below.

The Delphi method is roughly this: a questionnaire is drafted and iterated with panellists; experts then produce position statements replying to each question; these are circulated along with a summary of responses; and individual experts are allowed to adjust their own opinions over multiple rounds. The summary includes the median response and the interquartile range (between the 25th and 75th percentile). (The questions often have numerical answers, or can be represented numerically.) Experts whose answers are outside the interquartile range for a given item are asked to give arguments for their prediction. The process is then iterated a further two or three times (Cooke, 1999, pp. 12–14).

The Delphi method recognises that spontaneous concordance of judgement is unlikely. Its iterative discussions are aimed at generating consensus by investigating the sources of disagreement and eliminating them through debate and information sharing. There are a number of ways one might justify such a process. First, many Bayesians hold that the only explanation for disagreement is asymmetry of information (perhaps most famously Harsanyi, 1967; Harsanyi, 1968a,b). Delphi rounds aim to ensure that the relevant information is identified, shared, and discussed, so that all experts achieve a common understanding. Ideally, these experts should then come to hold the same beliefs. Second, Delphi rounds might be seen as emulating scientific discussion, as usually played out through a series of articles, peer review stages, and discussion in conferences. The conclusion of a Delphi process might then lay claim to something like scientific objectivity. I'm thinking of analyses of objectivity which take it to be the product of the right kind of intersubjective agreement, or which place certain social processes at the core of securing objectivity. E.g., Longino (1990, p. 62) claims that “the objectivity of science is secured by the social character of inquiry”.

However, the method faces a number of critiques. First, there are socio-psychological problems with “open” Delphi processes in which the participants’ identities are known. Aspinall (2010) argues that participants often revise their views on the direction of the supposedly “leading” experts, rather than in the direction of the strongest arguments. Longino (1990, pp. 76–79) anticipates this, noting that not all interaction increases objectivity; the following conditions are required: recognised avenues for criticism, shared standards that all can invoke, responsiveness to criticism, and *equal distribution of intellectual authority*. The final element is intended to exclude from “objectivity” those conclusions generated by communities in which “a set of assumptions dominates by virtue of the political power of its adherents.” Outside of the familiar guardrails of the institution of science (such as blind peer review) experts may feel unusual pressure to agree with big names on the panel, or to conform to a consensus in order to avoid becoming known as a “radical.”

Second, the method has an inbuilt disciplining function. “The respondents are not treated equally. People whose predictions fall inside the interquartile band are ‘rewarded’ with a reduced workload in returning the questionnaires, whereas those whose predictions fall outside this band are ‘punished’ and must produce arguments” (Cooke, 1999, p. 16). Or, in the words of Woudenberg (1991), “a Delphi is extremely efficient in obtaining consensus, but this consensus is not based on genuine agreement; rather, it is the result of. . . strong group pressure to conformity" (quoted in Morgan, 2014).

These lessons seem to generalise. If a consensus is to be built in a limited time, outliers must be shepherded towards the emerging consensus. If they are stubborn, some disciplining mechanism will be required. There may be less problematic ways of bringing this about than in the naïve Delphi process just described, but I suspect we will always find concerns about consensuses that are constructed under pressure. I am happy to accept that propositions which are subject to a genuine consensus have epistemic value, and though it is difficult to state exactly why it seems to be linked to the notability of multiple experts applying their minds, considering the evidence and arriving—of their own accord and individually—at the same conclusion. The introduction of time limits and disciplining mechanisms, and the prior stipulation that *some or other* collective position is the goal, undermine this.

Long-term consensus building is slow and politically fraught

The main vehicle for longer-term consensus-seeking is the scientific assessment. An assessment is a systematic attempt to review the state of expert knowledge on an issue, judge the quality of evidence, and advise on solutions to problems. They are convened by a governmental or intergovernmental body in order to develop understanding of, and support decision-making on, that issue. Their main product is an assessment report, which is a consensus document that is subject to peer-review by independent experts. In the 20th century assessments have been institutionalised as one of the major interfaces between government and researchers. They are now highly structured processes with vast resources, administered by government research bodies like the National Research Council in the USA. They are also extremely widespread, with the NRC producing 200-250 assessment reports per year (*About Our Expert Consensus Reports*). The organisations established to conduct assessments have long lives and significant impact on their research communities, with some operating for many decades and commanding budgets of hundreds of millions of dollars (Oppenheimer et al., 2019, Ch. 1).

Assessments collect and synthesise evidence, but they also create it. Assessments identify uncertainties and "gaps" in research that are particularly problematic for decision-making. In their role as research coordinators and funders, they then attempt to reduce this uncertainty by commissioning new research.

But, while providing resources and directing research can accelerate the pace of progress on a particular issue, this is no guarantee that uncertainty can be reduced on a convenient timeline. The large scale and significant bureaucracy involved introduce complexities that can retard progress. The establishment of a framework for gathering, assessing and integrating evidence is itself often contentious, leading to long debates in the lead up to the actual assessment. Integrating research from different disciplines and sources is time-consuming and difficult. The fact that assessments are convened by governments introduces an inevitable political element. Ideological factions vie for control of the report-writing, on the understanding that controlling what counts as "consensus science" is a means of controlling which policies are made. Government agencies and academic subdisciplines compete for control of research budgets. For example, Oppenheimer et al. (2019, Ch. 2) trace the competitive dynamics between the Department of Energy and the Environmental Protection Agency in the administration of resources during the first phase of the NAPAP, apparently as a means of enhancing the interests of their stakeholders.

Assessments like the IPCC's are so complex and time-consuming that they have been criticised as impeding rather than advancing research (Oppenheimer et al., 2007, p. 1506). In some cases, these difficulties are so severe that the assessments fail to provide their assessment reports in time for the relevant decision to be made. The NAPAP was established by the US government in 1980, but did not publish its integrated assessment report until 1991—one year *after* the passage of the 1990 Clean Air Act Amendments, widely regarded as the turning point in regulation on the issue (Oppenheimer et al., 2019, Ch. 2).

Focus on consensus impedes decision-making

Many (though by no means all) of the problems highlighted above are repercussions of the focus on consensus. But consensus itself may not be the right goal. Aspinall argues that "when scientists disagree, any attempt to impose agreement will 'promote confusion between consensus and certainty'." Agreement is merely a proxy for the epistemic goods we care about, and its elevation to a good in itself is dangerous. "The goal should be to quantify uncertainty, not remove it from the decision process" (Aspinall, 2010, p. 294).

The danger comes from the fact that consensus-seeking inevitably involves minimising results which are contentious. Michael Oppenheimer has been a leading champion of this issue. Oppenheimer et al. (2019, Ch. 4) discuss this at length for the case of the West Antarctic Ice Sheet (WAIS) assessments, and find that a focus on settled knowledge led to the omission of questions and results that later changed the consensus when fully appreciated. Much earlier, also in a discussion of the WAIS and IPCC assessments, Oppenheimer et al. (2007, p. 1505) reported that "setting aside or minimising the importance of key structural uncertainties in the underlying processes is a frequent outcome of the drive for consensus." In the context of policy decisions with much at stake, consensus is too conservative a standard and can lead to the dangerous underestimation of threats.

1.2 DECISION-MAKING WITH UNRESOLVED UNCERTAINTY

This thesis explores another option: decision-making in the face of unresolved uncertainty. The prior section establishes one reason we need such tools: to supplement flawed consensus-seeking decision-support processes. But regardless of whether the (admittedly brief) critique in the prior section is successful, such tools are clearly valuable. In some domains, there is no consensus and yet we cannot afford decision-paralysis.

This thesis considers three kinds of approach: Bayesian epistemology, opinion pooling, and ambiguity decision-theory.

To make philosophical progress, I restrict my discussion in various ways. I will focus on cases where the expert reports and model outputs involve probabilities. I take it that probabilistic opinions are of great interest as they form a core part of decision-making, both in theoretical studies of decision and in practical policy scenarios. Progress on how to resolve probabilistic disagreement is therefore extremely useful, although not fully general.

Probabilities are also central to formal epistemology and philosophical decision theory, where they represent agents' partial beliefs. My discussion here is framed in terms of the policymaking problem I am most interested in, but this work will have philosophical implications that go beyond this application. My decision theory and formal epistemology have comparativist foundations: I take agents to have two fundamental comparative attitudes that are relevant to my investigations: preferences and partial beliefs. Most of my focus will be on the latter. Partial belief is the name I will use for what is also called comparative confidence, or credibility, or comparative likelihood. (While the confidence language is common and intuitive, I will use the term "confidence" in a new way in chapter 5 and so wish to avoid confusion.) The link between comparative partial beliefs and probabilities is explored in the methodological reflection in chapter 7.

Methodology

A core criterion for the selection of a solution will be its practical applicability in cases like Ade's. I intend this work to be an exercise in non-ideal theory, or at least in less-ideal theory. To specify what that means, let me begin with what I take "ideal theory" to mean in the context of formal epistemology and decision theory. The labels "ideal theory" and "non-ideal theory" come from moral and political philosophy, where there is an ongoing debate about the fruitfulness of normative theorising in the presence of various idealisations. Mills (2005) provides a partial taxonomy of these, including ignoring certain characteristics of agents (such as power relations between them), ignoring cognitive limitations, and downplaying informational restrictions. Many of these "idealisations" also play a role in (formal) epistemology and decision theory. Formal epistemology constructs idealised models of agents and their doxastic situations, in a manner that is made precise in chapter 7. These models can be more or less idealised; that is, the agents in these models can have properties that are more or less distant from the properties of real agents like you and me. Examples of idealised properties include logical omniscience, instant computation, very rich priors, and so on.

Ideal theory is more than a set of assumptions, it is a methodology. As Mills puts it, "ideal theory either tacitly represents the actual as a simple deviation from the ideal, not worth theorizing in its own right, or claims that starting from the ideal is at least the best way of realizing it" (Mills, 2005, p. 168). In epistemological and decision theoretic ideal theory, it is thus legitimate to draw upon the idealised features discussed above when answering questions about what agents should believe or what they should do. For example, it is standard to assume that agents have prior probabilities over every proposition in an algebra, which itself can be very large and complex. It is no problem, when doing ideal theory, to recommend a complicated belief revision procedure, employing countably infinitely many probabilities concerning completely unrelated propositions.

Non-ideal theory attempts to hew closer to reality. It attempts to build a normative theory up from the messy ground-level reality, rather than down from the pure heights of ideality. In ethics, there are two traditions of non-ideal theory, according to Mills. Roughly, the first thinks that ideal theory is incomplete and needs supplementation; the second thinks that ideal theory is fundamentally misguided and needs to be replaced entirely. When I say that my project is "less-ideal theory" I mean to indicate that I am more aligned with the analogous first critique in the epistemology case. When doing non-ideal theory, the kinds of moves described in the paragraph above give us pause. At each stage, we ask ourselves whether the recommended procedure is within reach of real agents. We prefer models that are less idealised, procedures that are simpler. My interest is in formal tools, which inevitably involve some degree of idealisation. Nevertheless, I will attempt to remain on the "near" side of the spectrum of idealisation, and will make repeated reference to how the solutions I discuss could be implemented.

Another criterion for success will be respecting Ade's position as a layperson. Solutions that effectively require him to be an expert himself are not solutions, under my approach.

1.3 OUTLINE

I begin by taking the problem of expert disagreement in a policy setting as a question for epistemology. Chapter 2 discusses Bayesian approaches to expert testimony and disagreement, which were developed very much in the realm of ideal theory. I begin with a discussion of the orthodox Bayesian solution: supra-Bayesianism. In this view, an agent should respond to expert disagreement by conditioning on their evidence— i.e., the fact that the experts have each made the reports they did. I outline four problems for supra-Bayesianism: rational unawareness, cognitive burden, the (ir)relevance of priors, and its (in)sensitivity to testimony.

These motivate for a consideration of another popular Bayesian solution to single cases of expert testimony: expert deference. I begin with the traditional model of expert deference, as a constraint on the agent's priors. Expert deference as a constraint on priors does better on the latter two problems for supra-Bayesianism (irrelevance of priors and insensitivity to testimony), but trades them for two new problems: the arbitrariness of which reports are deferred to, and an inability to account for expert disagreement. I propose a new model, which I call expert deference as a belief revision schema, that deals with these problems. Expert reports are treated as exogenously given constraints on posteriors, and form part of a general belief revision schema. I discuss cases where the instances of that schema are belief revision rules such as Bayesian, Jeffrey and Adams conditioning. I also show how this model deals better with expert disagreement than alternatives. But "better" does not a solution make; significant issues remain and they motivate for consideration of other methods.

In chapter 3 I discuss the family of approaches to disagreement known as opinion pooling. This involves averaging the opinions provided by each expert; "averaging" can mean one of several mathematical functions. I discuss the two major pooling functions, linear and geometric pooling, in detail and comment on multiplicative pooling. I begin with the discussion of why one might think that linear averaging of opinions is any solution to disagreement. I survey a variety of arguments, from analogies with the statistical sampling to mathematical convergence results, and find all but one wanting. Linear averaging gains limited support from a mathematical result showing that in a particular circumstance choosing the linear average minimises one's expected error. I then discuss various characterisation results that show that if one wants a pooling function to meet certain plausible criteria, linear pooling is the only option. I conclude with some cautions against these weak motivations in favour of linear pooling.

My discussion of geometric pooling focuses on whether policymakers ought to be concerned with the so-called rational properties that such pooling functions exhibit. I argue that they ought not care about external Bayesianity or its variants.

I then turn to a discussion of a common ingredient to all pooling functions: weights that determine how much each opinion counts toward the aggregate. I argue that the selection of a scoring rule to determine these weights generates two problems. The first is a regress in which the policymaker faces a further expert disagreement. The second is a form of value-ladenness that closely mirrors the "argument from inductive risk" that has been extensively discussed in the philosophy of science. While these problems have solutions, together with the problems identified above they provide me with sufficient reason to look for a non-aggregative approach to expert disagreement.

At this point the thesis shifts from having a primarily epistemological approach to a decision theoretic one. It also shifts from focusing on expert opinions to focusing on the results generated by scientific models.

In chapter 4 I introduce scientific models, ensembles of models, and their role in decision support. I introduce a case study involving models of hurricanes in the North Atlantic. I revisit the arguments against averaging, and show that ensembles of models face the same problems. In addition, the nature of scientific models introduces a number of additional reasons to question averaging as a solution to disagreement.

I therefore turn in chapter 5 to a decision theoretic approach that does not involve any aggregation. I begin with a reconstruction of the “confidence” theory of decision-making under ambiguity. I rework it to take input from a model ensemble of the type discussed in the previous chapter. I apply the approach to the case study, showing how it handles insurance pricing decisions with the hurricane model ensemble. This demonstrates how the confidence approach gives policymakers a tool for making decisions with unresolved uncertainty *directly*— i.e. without selecting a single probability arbitrarily or aggregating to create one. It does not misrepresent uncertainty, as consensus-seeking and aggregative approaches do, but nor does it supply decision-makers with information about uncertainty that is useless or paralysing.

I then turn to a consideration of various ways of constructing the main ingredient in the approach: a nested family of sets of probabilities. I end with a discussion of various objections and concerns to the confidence approach. My conclusion is cautiously optimistic. As we currently lack good tools for making policy decisions with unresolved uncertainty, demonstrating the confidence approach's suitability to them is of value to policymakers and serves as motivation for philosophers to further study the approach.

This concludes the main part of my thesis and my study of the problem of policy decision-making and uncertainty.

In chapter 7 I turn to methodological and meta-philosophical reflection. In writing the preceding chapters, I had the opportunity to work in two often disconnected fields: decision theory/formal epistemology and the philosophy of scientific models. While chapters 4 and 5 involve applying tools from the former to a problem in the latter, chapter 7 does the reverse. Here I turn a philosophy of modelling eye on formal epistemology and decision theory. I argue that formal epistemology and decision theory engage in modelling of a kind that is very similar to scientific modelling. I do so by examining a particular model: a comparativist model of partial belief, and its connection with probabilistic models.

One crucial difference is that philosophy is often normative, while science is typically not. I examine the difference this makes, and conclude that much of the theory of modelling can fruitfully be applied to normative models in philosophy. This allows us to draw methodological insights from the practice of scientific modelling and use them to inform our philosophical practice. In particular, I argue that the idealised and indirect nature of modelling is under-appreciated by philosophers. A number of inference-patterns familiar from other parts of philosophy do not work well in a modelling context, including certain realist inferences, and reasoning by counterexample. My discussion also casts new light on familiar debates about representation theorems, and the dispute between precise and imprecise probabilists.

Notation and modelling preliminaries

This work makes repeated use of mathematical models of agents' beliefs. I will here provide some basic introductions to the notation and concepts used in the formal work to come.

The expert disagreement problems I am interested in involve one agent (the policymaker, who is a novice) receiving reports of the opinions of a number of others (the experts). These "opinions" will be the experts' beliefs about various events that are of interest to the policymaker. Specifically, the experts will report the probabilities for those events, expressing the experts' degrees of belief.

I will represent agents' opinions by functions, typically denoted P for the policymaker and P^i for the experts, where i is an index running from 1 to n, which is the total number of experts. These functions are initially taken to be probability functions defined on a Boolean algebra of propositions, though this assumption is later relaxed.

I call the agent's attitude *partial belief*, and occasionally refer (as I did above) to their degrees of belief. The function representing that attitude is called a credence function. "Credences" is a general term for the mathematical avatars of the agent's partial beliefs; "precise credences" are numbers generated by a probability function, while "imprecise credences" are sets of numbers generated by sets of probability functions. In each case, those numbers (or sets of numbers) represent the agent's partial beliefs.

Propositions are denoted by capital letters from the end of the alphabet, e.g., X, Y. A "Boolean algebra" is an algebraic structure: a set of propositions, equipped with the logical operations $\wedge, \vee$, and $\neg$ which are interpreted as logical conjunction, disjunction and negation, obeying the usual rules. The set is closed under negation, and under finite disjunctions or conjunctions of propositions. So for any X, Y in the set, $\neg X$ and $\neg Y$ are in the set, as are $X \vee Y$ and $X \wedge Y$. These logical operations partially order the set, thereby giving the structure a top element $\top$ and a bottom element $\bot$. These are also called the tautology and contradiction, respectively, and $X \vee \neg X = \top$ and $X \wedge \neg X = \bot$ for any proposition X in the algebra. Propositions that are not equivalent to either $\top$ or $\bot$ are called "contingent." I make occasional use of partitions: these are sets of propositions $\mathbb{X} = X_1, \ldots, X_m$ such that $\bigvee_{i=1}^{m} X_i = \top$ and $X_i \wedge X_j = \bot$ for $i \neq j$.

The experts have a common domain of expertise, which we will model as an algebra denoted Ω, from which the propositions we consider are drawn. Let $P : \Omega \rightarrow \Re$, then if $X, Y \in \Omega$, P is a probability function iff $P(X) \geq 0, P(\top) = 1, P(X \vee Y) = P(X) + P(Y)$ if $X \wedge Y = \bot$. For simplicity, I will consider only finitely many propositions and assume that all these probability functions are regular, i.e., that they assign non-zero probability to any contingent proposition.

At various points, I will talk about evidence for propositions. The agent's evidence is modelled by a proposition, typically denoted E, which represents the logically strongest proposition the agent knows (in the case of current total evidence) or that they learn (in the case of new evidence).

Upon receiving some evidence, including the testimony of an expert, an agent revises her opinions. Her new probabilities are denoted Q or Q^i if she is one of the n experts. I will often consider the experts' opinions from the perspective of the novice. We will want our novice to respond to reports of the form "the probability of X is x^i" by expert i, and so I will consider propositions that capture such facts. (x^i will always be the probability assigned to X by expert i.) I use the notation $\ulcorner P^i(X) = x^i \urcorner$ to denote the proposition that expert i reports probability x^i for event X.[6] What the novice learns is this proposition, $\ulcorner P^i(X) = x^i \urcorner$. For brevity I will occasionally write $\ulcorner P^i \urcorner$ when highlighting the reporter's identity, or $\ulcorner x^i \urcorner$ when highlighting the content.

[6] I use the term "event" in place of "proposition" in situations where it is helpful to distinguish propositions concerning reports of expert's probabilities from propositions concerning other things—the latter being "events."

2

EXPERT DEFERENCE AND DISAGREEMENT

2.1 INTRODUCTION

This chapter presents a broadly Bayesian analysis of expert testimony and disagreement. Bayesianism is what passes for a benchmark theory of rationality in decision theory and formal epistemology, and so is a natural place to begin. In line with the approach outlined above, I will examine how Bayesian tools function for realistic agents in policymaking scenarios, seeking to preserve what is attractive about the theory while coming closer to the capabilities of real agents.

At its core, Bayesianism is committed to two norms: that one ought to have probabilistic partial beliefs, and that one ought to update those beliefs by conditioning on one's evidence.

How reasonable the theory is—for real or ideal agents—depends in part on how we interpret these norms. Often, they are assumed to be *evaluative norms*: they are features of a good believer. Evaluative norms needn't entail anything about action: a good spring day is cloudless and fresh; these are evaluative standards for assessing days *qua* spring days, but do not directly bear on the actions of any agents. But the Bayesian norms are also sometimes taken to be *action-guiding*. This is especially so in the Bayesian statistics literature where investigations of, for example, expert testimony, include discussions of how real agents might carry out Bayesian processes.[1]

I am interested in policymakers facing actual cases of expert disagreements, and my aim is to contribute to advancing their practice. Therefore, I am interested in guidance for action. There is, of course, a link between evaluative and prescriptive norms. Evaluative norms can give rise to prescriptive ones: rules for baking bread are created with good bread as their target. In the other direction, facts about what one *can do* may constrain standards for evaluating one's goodness. If the prescriptions associated with an evaluative standard are impossible, this may require a revision of that standard.

I begin with these preliminary remarks on the nature of Bayesian normativity in order to have the issue in view as I discuss Bayesian approaches to expert disagreement, and criticisms thereof.

I begin my discussion with the most orthodox Bayesian approach, supra-Bayesianism, which is often presented as being identical to consensus building in idealised conditions. I argue in section 2.2 that it is not, in fact, a credible solution. I introduce four problems for supra-Bayesianism, which then act as goals for further approaches. There is a long history of discussions of supra-Bayesianism, so this section is largely a presentation of existing arguments.[2]

[1] For an explicit discussion of these two kinds of norms in epistemology see (Simion, Kelp, and Ghijsen, 2016, S4.1), and for a similar discussion in decision theory see (Buchak, 2013, Ch 1) and (Thoma, 2019). The Bayesian statistics papers referenced in this section almost all have a prescriptive element, but for a particularly clear example see (French, 1980).

[2] This section, and indeed this chapter, is not intended as a complete survey of the literature on Bayesian approaches to expert disagreement and, where I do survey the literature, my review is partial to philosophy. There is a Bayesian statistics literature on the topic of expert testimony covering both supra-Bayesianism and deference, and I engage with it here only partially. Part of the difficulty in using that literature arises from the difference in focus. Statistics papers often *assume* that orthodox Bayesianism

In section 2.3 I begin my discussion of expert deference principles, which are common starting points for formal discussions of expertise. I start with the traditional way these principles have been presented, as a constraint on priors for an ideal Bayesian agent. I argue that they still fall prey to the problems I introduced for supra-Bayesianism, and therefore are insufficient solutions to the problem of expert testimony. They fare worse still as a solution to expert *disagreement* (i.e., multiple, contrary testimonies), as I go on to argue.

The remaining sections depart from orthodoxy. They are "Bayesian" in the sense that they involve probabilistic models for partial belief, and that they are committed to belief revision by Bayesian conditioning in *some* cases. But they entertain an increasingly wider set of belief revision procedures and non-standard methods.

In section 2.4 I present a new development of the expert deference idea, in a manner that (a) is better suited to my my non-ideal theory approach, and (b) can accommodate expert disagreement. I call the new approach "expert deference as a belief revision schema," and draw on the theory of probabilistic belief revision in order to develop it. The new approach can handle a much wider class of expert reports than the two orthodox models. I show that this new approach does not fall prey to the four problems that bedevil supra-Bayesianism and expert deference as a constraint on priors.

is the right norm, while I wish to evaluate that claim. They work through how a real agent might reason in the kinds of cases under consideration, and regularly assume a particular form for the agents' priors and likelihoods (i.e., assuming particular distributions) in order to make progress. While valuable for building understanding of Bayesianism and its implications, they are rarely directly concerned with my topic here.

I then turn in section 2.5 to the problem of unawareness. I show that my new expert deference proposal fits neatly into models of awareness growth, and therefore can function as part of a complete heterodox Bayesian solution to expert testimony.

In section 2.7 I consider an alternative belief revision process to those discussed in section 2.4—revision by divergence-minimisation. I present a methodological argument for why I do not think this strategy is a fruitful way of tackling the general problem of expert testimony.

2.2 SUPRA-BAYESIANISM

To begin, let us consider an example where an agent receives testimony from just one expert.

Example 1. *You open your weather app and see, to your complete surprise, that there is a 30% chance that London will be struck by a hurricane on Thursday.*

The basic idea that I will be exploring is that the fact that this information comes from an *expert* means that the content of their report, that there is a 30% probability of a hurricane in London on Thursday, ought to *directly* influence your credence in that proposition. The question this chapter looks at is then: how do we model this in a framework where agents' partial beliefs are represented probabilistically?

To start, I'll look at two Bayesian answers to the question and draw out some problems with them. These problems will then act as targets for a new model.

The first answer is often called supra-Bayesianism in the context of expert disagreement (many experts giving contrary testimony), and I will use it here even in the one-expert case. It is simple Bayesian orthodoxy: when you hear the expert report, you update your beliefs by conditioning on what you have learned.

The setup is standard: your probabilities P include a prior for H, the proposition that a hurricane will hit London on Thursday, and when you look at your app you learn the proposition that the weather app says "there is a 30% chance of a hurricane", denoted $\ulcorner W(H) = 0.3 \urcorner$. Upon learning that $\ulcorner W \urcorner$, your posterior probability for H is:

$$Q(H) = P(H | \ulcorner W(H) = 0.3 \urcorner) = \frac{P(\ulcorner W \urcorner | H)}{P(\ulcorner W \urcorner)} P(H)$$

So, your posterior for H depends on your prior for H, your prior for hearing this report $\ulcorner W \urcorner$, and your prior likelihood for hearing the report, given that there will be a hurricane.

This answer, though perfect Bayesian orthodoxy, raises a number of worries.[3](These are most acute if, like me, you are interested in applying your Bayesian theory to real cases of expert disagreement, for which Case 1 is an exemplar, but they also concern anyone interested in a "realistic" theory of rationality for bounded agents.)

[3] The label "supra-Bayesianism" comes from Keeney and Raiffa (1976). It has been much discussed in the Bayesian statistics literature (see Genest and Zidek, 1986), and I do not claim that these problems are without possible responses. In particular,

1. *Awareness:* In the setup of the example, I stressed your surprise at hearing this report. London doesn't get hurricanes, and so it is natural to say that you'd never considered H before, much less the proposition that the weather app would report precisely 30% as the probability for it! It is implausible that you have any views on these matters at all, yet the supra-Bayesian view insists that you have priors for them.

When I talk of "awareness" I do not mean that agents must actively reflect on something, or that they pay attention to it. Put simply, if an agent has never thought of something, they're unaware of it. An extreme case of unawareness would be an agent's relation to a proposition involving a concept that the agent does not have—perhaps, your present relation to some esoteric statement about theoretical particle physics. Or consider Whorf's discredited claim that the Hopi lack our concept of time. Utterances involving such concepts are effectively opaque to the agent; upon hearing them, they do not apprehend the proposition the speaker intends to express, nor have they ever done so. Between these two extremes—mere inattention and conceptual lack—lie many cases where the agent does not meet common criteria for belief. They may lack any relevant disposition to act, or stand no relation to the relevant mental representation. If you know nothing about South African politics, then you know nothing about the Democratic Alliance or their electoral hopes in the city of Johannes-

much work has been done on how to make it more tractable in cases where particular symmetries, or known distributions, simplify the updating required. Lindley (1982) notes cases in which it reduces to the very simple expert deference. Others have studied when it reduces to averaging. French (1980) is an early analysis of how thinking through the procedure a real agent might use to enact supra-Bayesianism can generate plausible simplifications.

burg. When confronted for the first time with my claim that the Democratic Alliance has no future in Johannesburg, you are in a state of unawareness.

Nevertheless, supra-Bayesianism requires agents to have attitudes towards all propositions. And not just upon hearing them; you are required to have had an attitude about the Democratic Alliance before hearing of it, just as you are required to have a prior belief about hurricanes in London on Thursday. Indeed, not only must agents have partial beliefs towards propositions like $\ulcorner W \urcorner$, but (partial beliefs representable as) precise credences.

To stress the implausibility of this, let me point out that that in order to apply supra-Bayesianism generally we must require that you have credences for *any* report (i.e., any $x \in [0,1]$) on *any* proposition made by *any* expert.[4] If, like me, you think that there are many combinations of expert, proposition, and report, toward which even the most rational agents will have no attitudes, then you think Awareness is a problem for supra-Bayesianism.

2. *Cognitive Burden:* There are a great many experts in the world, a myriad of propositions they might report on, and a continuum of reports they could make on each. Experts and their reports may have complex dependencies on one another. Ω must therefore be a very rich algebra indeed, and the range and granularity of the judgements the agent is required to make are breathtaking. Any procedure to

[4] One might also worry that this demand, taken literally, means that the simple model above won't work. Experts report probability values, and so these reports are themselves continuous random variables. Strictly speaking your prior for $\ulcorner W(H) = x \urcorner$ should thus be zero, for any x. I won't dwell on this problem, as the issues it raises aren't core criticisms of supra-Bayesianism and I believe that a more complex model could work around it.

enact (even approximately) the supra-Bayesian answer is therefore extremely cognitively demanding for any real agent (this has been extensively discussed; see e.g., the comments on supra-Bayesianism in Genest and Zidek (1986)). It is no failing of rationality not to be able to accomplish this procedure.

This objection concerns what real agents can achieve. A Bayesian might quibble: a rational agent simply does behave in a manner that is (representable as) complying with the diktats of Bayesianism. In this case, that involves supra-Bayesian updating on the reports of any experts giving testimony. But it is no more assumed that agents actually perform these calculations than it is assumed that one does mental trigonometry when catching a ball. Catching a ball *amounts to* calculating a trajectory and performing a sequence of movements such that one's hands intersect with that trajectory, but that statement can be true independently of what is going on in the mind of the agent doing the catching.

I acknowledged that there is no requirement that agents be consciously attending to the beliefs we are discussing, or consciously perform any calculations. The Bayesian can say that a rational agent has cognitive architecture that accomplishes the belief changes prescribed by Bayesianism *somehow*. Perhaps if we supplement this with a "low cost" analysis of belief such as dispositionalism, then the Bayesian can insist that there aren't direct cognitive demands on agents. As long as they end up with the right dispositions, they're rational. The costliness of performing the calculations to determine the right posterior belief in the representation can completely decouple from the costliness of accomplishing the actual belief change in the agent's cognitive architecture.

Depending on the cost, this may be a reasonable response if one's aim is to construct a theory of rationality that interprets the rational behaviour of real agents; or one that is normative in the sense of providing an evaluative standard only. But if one takes Bayesianism to be action-guiding, then it *is* a flaw if no real agent could reproduce the obvious procedure for determining which are the right actions (i.e., permissible posterior beliefs) and if no alternative procedure for that determination is provided. As I am interested in generating normative guidance for policymakers facing expert disagreement, I view it as such a problem.

3. *Relevance of Priors:* One might reasonably ask: what do you know about hurricanes, anyway? The reliance of $Q(H)$ on $P(H)$ strikes many as problematic: surely it is rational to jettison your ignorant prior in face of reliable expert testimony? Similarly, why should $Q(H)$ depend on how likely you think this expert is to report precisely 30% as their probability for H? What do you know about hurricane prediction, or the methods of this or that forecaster?

4. *Sensitivity to Testimony:* The complementary problem to supra-Bayesianism's over-sensitivity to your priors is that it is *under*-sensitive to the actual content of the expert's report. $Q(H)$ isn't a function of the expert's reported credence! It is instead a function of your priors that the expert will *report*, in this case, 30%. This is because the supra-Bayesian procedure is just the general Bayesian answer to every learning experience. But it seems like there's something different

going on in the case of expert testimony: you're receiving information that is *directly* relevant to your credence in H, in a way that is unmediated by your credences in the learned proposition.

To this the Bayesian can reply the generality of the approach is a strength, and that in this case that strength manifests as a sensitivity to the fact that a particular expert made this report in the precise manner that they did. This complex fact is the right thing for the agent to respond to, as it allows their response to depend on what they think about this expert, what they know about the circumstances of the report, and so forth. The precise value reported is just one of many features of the learning experience to which the agent should be responsive.

This is simply a doubling down response. It denies that there is anything special about expert testimony. But there does seem to be something particular about a case of expert testimony: this learning experience contains as one feature of it a number that our credence ought to be close to, assuming that certain requirements are met.

Taken together I regard these worries as providing sufficient cause for concern about supra-Bayesianism that I wish to find an alternative. As outlined above, I am concerned about real agents; in particular their ability to extract action guidance or a comprehensible normative standard in a variety of cases. I will therefore consider alternatives that are more limited than supra-Bayesianism, but easier to enact.

2.3 EXPERT DEFERENCE AS A CONSTRAINT ON PRIORS

One intuitive thought about expert testimony is that, under the right conditions, laypeople should defer to it. Deference means adopting the expert's testimony into your beliefs. For example if you are interested in whether it will rain, you should ask the weather forecaster and believe what they tell you: if they say the chance of rain is 40%, you should believe that and thus set your own credence in rain to 40%.

As I am interested in probabilistic opinions, I will think of deference in this way, as taking an expert's probabilities on as your own. Deference is so common as a thought about expert testimony that some have taken it to be the definition of an expert: Gaifman (1988, p. 193) defines an expert as someone for whom "the mere knowledge of... [their] assignment will make the agent adopt it as his subjective probability." This definition is common in Bayesian statistics (for a contemporaneous usage see DeGroot, 1988), and has been adopted in philosophy by, e.g., Joyce (2007) and Elga (2007).[5] This definition is somewhat unhelpful if one is interested in *identifying* which people are experts, but it does highlight the centrality of the deference idea.

Deference may strike the reader as a rather extreme idea. Would a real person defer to a real expert? Arguably we do so all the time. The presumption that people speak truly in ordinary conversational

[5] There are alternate definitions of expert out there. For example, Easwaran et al. (2016) define experts as reliable witnesses. For them, P_1 is an expert for P, in some domain D, when the following holds: for any $X \in D$, when $P_1(X) > P(X)$, P takes P_1's credence in X as evidence for X and raises their credence. The same applies to lower credences as evidence against.

contexts results in something like deference: taking propositions to be *true* because someone reports them to be true. The practice of learning science involves accepting the material in textbooks in a manner that is close to deference. Scientists do this too: experimenters working with radioactive materials do not assess molecular half-lives for themselves, they look them up. More prosaically, I often look at the weather forecast and act accordingly. Later, if someone asks me whether it will rain, I often quote the reported chance of rain.

Now in this latter case I may not *precisely* defer: weather forecasts are famously ridiculed for their inaccuracy, and many take them with a pinch of salt. We might think of this as "partial deference", perhaps defined as a form of averaging—updating to a mixture between my prior probability and the expert's report: $Q(X) = \alpha P(X) + (1 - \alpha)P^i(X)$ for some $0 < \alpha < 1$ (e.g., Joyce, 2007, pp. 190–91). Consideration of partial deference, so defined, can therefore be delayed until chapter 3, when I turn to opinion pooling. That chapter also includes a discussion of how one might decide on the relative weights to assign to different experts. In this regard it is an improvement on the supra-Bayesian position, which insists that the weights given to experts are bundled into the Bayesian update in the likelihoods that the agent subjectively assigns to hearing each particular report from that expert. The supra-Bayesian way of dealing with the reliability of experts is therefore subject to the Relevance of Priors objection, while the less subjective procedures discussed in chapter 3 will seek to avoid them.

Regardless of whether we *do* defer, one might reasonably worry whether we ought to. One way to see the idealisation involved in deference is to note its link to calibration. Consider a case where the relevant reports concern something that occurs multiple times.[6] A probabilistic report is called *calibrated* when propositions X assessed to have probability $x\%$ turn out to be true $x\%$ of the time. Put another way, if we collect all of the predictions that something is $x\%$ probable then, if those predictions are calibrated, the proportion of events that turn out to be true will be x.

Calibration is a statement linking the report $\ulcorner P^i(X) = x^i \urcorner$ with the actual frequency of occurrence. If we know a predictor is calibrated, then we can project these frequencies into the future—calibrated predictions are *chance signals*. If our agent obeys a principle that known chances should guide her credences, then she ought to set her posterior degrees of belief to match the calibrated prediction.[7]When we defer to *uncalibrated* experts, the idealisation we're making is to treat them as if they were calibrated. But again, this can be reasonable: experts are often the best stand-in that we have for chance signals, and we should still expect to do better by deferring than we would by holding to our prior credences. While we know that they aren't perfectly calibrated, we don't know the way in which they fail to match the frequencies. Deference is a simple, easy to implement, procedure to improve our own predictive performance.

[6] We may of course wish to defer to experts on matters which occur only once, in which case this notion of calibration to frequencies isn't useful.

[7] Lindley (1982, p. 118) notes the connection between calibration and deference in an early discussion of supra-Bayesianism and deference. When the result of supra-Bayesian updating matches the expert's report, Lindley calls the expert "probability calibrated" for that novice. Given my earlier discussion of the Gaifman/DeGroot definition of expert, I won't use Lindley's terminology for fear of confusion. As De-

The classic way to model expert deference is as a set of priors that the agent has. Recall Example 1, where you look at your weather app and it states that there is a 30% chance that a hurricane will strike London on Thursday. We can model you as deferring to this expert by equipping you with what I will call a "deference prior", a set of prior credences defined by $P(H|\ulcorner W(H) = x\urcorner) = x$. This covers all experts W, all propositions they may report on (in their domain) H, and all reported values for their credence $x \in [0, 1]$. So when you hear the report in Ex.1, and learn the proposition $\ulcorner W\urcorner$, you update by conditioning and get:

$$Q(H) = P(H|\ulcorner W(H) = 0.3\urcorner) = 0.3$$

We can immediately note some pros to this approach. The idealisation that you simply adopt the expert's reported credence has the benefit that the model is much simpler than supra-Bayesianism. It is Sensitive to Testimony by construction, as the answer depends directly on the content of the report. The problem of the Relevance of Priors is alleviated: it doesn't use your prior for H or $\ulcorner W\urcorner$ to arrive at your posterior for H, but instead fixes your prior for H conditional on $\ulcorner W\urcorner$ (thereby constraining your priors for H and $\ulcorner W\urcorner$, but these don't play a direct role).

The problem of Cognitive Burden is reduced, if we think an equation fixing a whole set of priors is less demanding than having free priors for each W, H and x. This doesn't resolve the part of the Cogni-

Groot (1988, p. 299) says: "it would be unnecessary to use the term 'well calibrated' in this paper because that property is now simply the defining characteristic of an expert."

tive Burden problem that is directed at the granularity and sensitivity of your prior attitudes, however. At least not if these priors are all interpreted as fully-fledged partial beliefs. A defender might respond that priors *aren't* truly beliefs, they are a representation of the agent's evidential standards (see, e.g., Titelbaum, forthcoming, Ch. 4). (Presumably, this view is primarily about prior conditional probabilities.) They tell us how the agent is disposed to respond to various pieces of evidence they may learn. So when we include in our model a constraint like $P(X|\ulcorner P^i \urcorner) = x^i$, we aren't imputing any *attitude* to the agent at the time when P is their credence function. Rather we're saying something about how they will respond when they learn that the proposition $\ulcorner P^i \urcorner$ is true. A very general principle ("defer to experts") can cover a great many propositions without requiring that the agent take any attitudes before receiving that evidence. I will return to this in the next section, as this is close to the strategy that I explore there—though the emphasis is subtly different.

More pressingly, the problem of Awareness remains in full force: the fact that the agent has attitudes to these propositions is implausible and has no part in the requirements of rationality. We might grant the defender of deference that the agent has an evidential standard that guarantees them an attitude to the proposition the expert directly reports (i.e., that there will be a hurricane on Thursday). But that is an attitude to only one in a complex web of propositions. Upon becoming aware of proposition H—and assigning it a worryingly high probability—the agent ought to update their other beliefs, e.g., that their car parked outside is at risk, given that there will be a hurricane. The classical model of deference appears to insist that the agent *already had* attitudes connecting H to all others. But to do so is to already have been aware of H, and to have attitudes conditional upon it. I deny that this is plausible.

Expert deference as presented also faces two unique problems. First, Arbitrariness: it isolates one kind of expert report as worthy of deference above others. Experts reliably report all kinds of things, and yet expert deference is a principle about deferring to reports of unconditional probability. Yet experts can (and do) report all manner of probabilistic information[8] including conditional probabilities, Bayes factors, comparisons between the probabilities of various events and variables, expected values for variables, functional forms for distributions over variables, and so on. It is unclear how a simple deference principle, which insists on operating through the Bayesian belief revision process, can capture all of these. But what reason could we have for distinguishing only reports of unconditional probability as worthy of deference?

Second, Expert Disagreement: it is unclear how to extend this expert deference principle to the case of more than one expert, when they disagree. Supra-Bayesianism's answer to this is the same as ever: conditionalise! One merely updates on the evidence received, making use of the relevant likelihoods and priors for each expert's report. This has all the problems discussed above for the one-expert case, but it *is* an answer. Expert deference doesn't seem to provide much of an answer at all. One cannot simultaneously defer to two experts (as I

[8] Experts report non-probabilistic information too, but here I'll neglect such reports. We can perhaps assume, as many probabilists do, that categorical statements (e.g., "it will rain tomorrow") are expressions of high credence ($P(\text{rain tomorrow}) \approx 1$).

have set things up, that operation simply isn't defined), and if one defers in sequence then the last report will dominate.

Deference treats all experts equally; as though they were the *same* expert. And if one received two reports from the same expert on the same topic, and deferred to them each time (assuming perhaps that they'd learned relevant new information), then of course the latter report would dominate. (This is desirable in the one-expert case. Suppose we start thinking the chance of rain is 20%. An expert says it is 40% and so we revise accordingly. The same expert later says it is 20%. Assuming that we maintain our view of them as an expert, we want to end up in the same belief state we started in: a credence on 0.2 in rain. The current setup gets us this.)

But it also reveals that our model has oversimplified by not distinguishing between the following: (a) a report from one expert, and unanimous reports from many, and (b) a sequence of different reports from one expert, and a profile of disagreeing reports from many. In what follows, I will be concerned at different times with each of these distinctions.

2.4 EXPERT DEFERENCE AS A BELIEF REVISION SCHEMA

My own proposal is a development of the expert deference idea, but one which hopes to avoid the remaining issues discussed above. In order to motivate for it, let us examine the source of our continuing problems with expert testimony. At the heart of the matter is the proposition $\ulcorner W(H) = x \urcorner$, "that the expert W reported x as their probability for H," which is what these Bayesian models take the agent to learn when they hear the report. The problems discussed above arise from:

1. having a proposition in the algebra to represent $\ulcorner W \urcorner$ and (all of the) H's—in the discussion above, this was linked to the problems of Awareness, Cognitive Burden, and Relevance of Priors.
2. having priors for (and related to) each value of x for each W and H—Awareness, Cognitive Burden, Relevance of Priors.
3. updating by conditioning on the fact of the report, rather than using its content directly to change your credence—Sensitivity to Testimony.
4. failing to provide a mechanism for delineating and dealing with multiple expert reports—Expert Disagreement.

In developing a better model, I propose to address each in turn. First, I will remove $\ulcorner W \urcorner$ from the model entirely. Second, I will allow for the fact that agents aren't aware of propositions like H, and so don't have priors for them or reports about them. Third, I will use a different belief revision strategy that is sensitive to the content of the report. It will turn out to provide at least some traction on the problem of expert disagreement, though by no means a full solution.

In brief, I propose to regard expert deference as a *belief revision schema*. By this, I mean that I will take expert reports as prompts to revise one's credences in a manner that "fits" the content of the report. To make a start, I will set aside the problem of Awareness and develop the model for familiar propositions. In the following section, I will expand the model to cover novel propositions.

In developing this model, I take inspiration from the literature on Jeffrey conditioning. This is the name given to a rule developed by Richard Jeffrey to deal with circumstances like this one.

> *The agent inspects a piece of cloth by candlelight, and gets the impression that it is green, although he concedes that it might be blue or even (but very improbably) violet. (Jeffrey, 1983, p. 165)*

Let G, B and V stand for the propositions that the cloth is green, blue or violet. Suppose that these form a partition for the agent, and that after their inspection to agent comes to have posterior degrees of belief 0.7, 0.25, and 0.05 respectively.

What has the agent learned here and how does it affect their beliefs? Jeffrey says: if there were a proposition E in the domain of the agent's credences describing the precise quality of this experience, then we would simply say "the agent learned E." The rational response to this learning experience would then be to update their degrees of belief by conditioning on E—this ought to be how they arrived at the posteriors for G, B and V, and how they update all their other beliefs. But, says Jeffrey, there needn't be any such a proposition in their algebra, nor expressible in English. Anything we could specify would be too vague to convey the precise quality of this uncertain experience, and too vague to support any meaningful ascription of precise conditional probabilities as the Bayesian procedure requires. It is better to admit that there is *nothing* the agent learns for certain.[9]

Instead Jeffrey proposes that we *describe the effects* of the experience on the agent, by stipulating their credences after the experience. We can omit the proposition E altogether, and merely say that they come to have degrees of belief for the partition $\mathbb{A} = \{G, B, V\}$ equal to $(0.7, 0.25, 0.05) = (\pi_A)$. Jeffrey then provides us with a rule for generating a fully-specified, unique and coherent posterior credence, now called Jeffrey conditioning.

Definition. (Jeffrey conditioning) Suppose an agent with initial probability function P comes to have new probabilities for a partition $\mathbb{X}$, denoted π_X for each X in $\mathbb{X}$. The agent's new probability function Q is obtained from P by Jeffrey conditioning if and only if, for all $Y \in \Omega$

$$Q(Y) = \sum_{X \in \mathbb{X}} P(Y|X)\pi_X. \tag{1}$$

A point we will come back to: Jeffrey argues that this is the right rule for revising belief whenever the agent's conditional beliefs given $\mathbb{X}$ remain unchanged; that is, for all $X \in \mathbb{X}$, $Q(\cdot|X) = P(\cdot|X)$. This is called the *Rigidity* condition. Indeed, Eq. 1 is just the law of total probability for Q when Rigidity holds.

Here is the point I want to take away from this. An orthodox Bayesian *can* model this situation. They simply insist that there *is* such a proposition E, capturing the precise quality of this experience. (Perhaps it is a "sense data" proposition, inexpressible in any natu-

[9] This is linked to Jeffrey's rejection of what he calls "dogmatic empiricism", the view under which there is some basic, sense-data proposition capturing exactly what the agent learns.

ral language but corresponding to the precise nature of the agent's experience.) They continue to insist that the proposition *is* a part of the agent's algebra, they have priors for it, and when they learn it they update by Bayesian conditioning. Jeffrey's approach represents a different *modelling strategy* for the same problem. He removes the unrealistic data propositions E from the model, and instead takes the experience to provide an exogenous constraint on the agent's posterior credences. He then proves that, for this kind of constraint (unconditional probabilities over a partition) there is a unique "kinematic" update rule that fixes a posterior credence function Q (Jeffrey, 1983, pp. 164–68).

I propose to do the same: take expert testimony to provide exogenous constraints on credences, rather than modelling it endogenously.

2.4.1 *First pass*

I'll start by sketching how this would work for an easy case, where the agent is aware of the proposition the expert reports on. Consider example 2.

Example 2. *You open your weather app and see that there is a 60% chance of rain this evening.*

Here we can assume that you are aware of the possibility of rain (you live in London, after all) and so we assume there is a proposition $R \in \Omega$ representing it. We needn't assume that you have a precise prior for it: we can say your comparative beliefs were incomplete with respect to R, or that you had completely imprecise credences for it, so $\mathbb{P}(R) = [0, 1]$. As we will see, the details won't matter for this first pass.

I propose that we model the expert report as providing a constraint that $Q(R) = 0.6$. In the ideal case of full deference, this is a precise constraint. There is no proposition in the model representing the expert's report, and we don't model the agent (you) as coming to learn any proposition for sure. Instead, your posterior is bound to obey this constraint.

Now typically our beliefs are multiply connected and other beliefs will depend upon this one. If I change my credence in rain, I will also change my credence in having an enjoyable cycle in to work, and my credence in arriving dry if I walk, and so on. If I fail to do so here, my credences will be incoherent. So some revision is required for the rest of my degrees of belief. Put in terms of our model, the probabilities of various other propositions $(Y, Z, \ldots)$ that are probabilistically dependent on R ought to change when the probability of R changes.

My proposal is to generate the remainder of the posterior credence by Jeffrey conditioning on the partition $\{R, \neg R\}$.[10] So, for example, letting "late" stand for the proposition that I walk to work and arrive late:

$$Q(\text{late}) = P(\text{late}|R)Q(R) + P(\text{late}|\neg R)Q(\neg R)$$

The language I just used is procedural, reflecting my interest in what real agents might do when confronted with expert testimony.

[10] This suggestion is due to Steele (2012).

First, one *fixes* the credence R. Then, one *generates* the relevant posterior credences by Jeffrey conditioning. This is explicitly envisaged as a process an agent might follow.[11] This assumes, of course, that they are in possession of numerical probabilities for the relevant propositions. That won't normally be the case for prosaic matters like being late to work. But in policy situations these probabilities are often estimated. A policymaker's ability to follow the procedures discussed in this chapter will therefore depend on which probabilities are available to them.

This is our first look at what I call *expert deference as belief revision*.

2.4.2 *Belief revision theory*

I will now introduce some ideas from the theory of belief revision to make this more precise. Following the formalism of Dietrich, List, and Bradley (2016), I will think of a belief revision rule as a function, mapping an initial belief state and an experience to a final belief state. Let $\mathcal{P}$ be a set of possible belief states, $\mathcal{I}$ be a set of possible inputs or experiences, so that a belief revision rule maps $\mathcal{P} \times \mathcal{I} \to \mathcal{P}$. Belief states will be probability functions, or sets of probability functions, as before. "Inputs" are taken to be very general, including straightforward observations, noisy signals, and expert reports of various kinds. We therefore specify them *extensionally*, as the set of belief states that are consistent with the experience. A simple example: if I look out the window and see that it is cloudy, this input constrains my belief

[11] In this I follow Jeffrey himself (e.g., Jeffrey and Hendrickson, 1989) and much of the Bayesian statistics literature.

state to include only those in which it is cloudy outside my window. We have just seen this idea at work in Jeffrey updating.

Belief revision rules can be characterised by two conditions: Responsiveness and Conservatism. Loosely, Responsiveness ensures that the final belief state "respects" the input, and Conservatism ensures that the belief revision changes *only* what is "required" by the input. This is captured by a Conservation condition, that specifies which parts of the prior belief state must be conserved by the revision.

Rules which follow these two conditions follow a pattern called perturbation-propagation.[12] First, as the rule is *Responsive* to the experience, the input will directly bring about a change in belief state: the perturbation. Second, the remainder of the belief state is adjusted to reflect the impact of the input; this makes use of the perturbation and the parts of the initial state that are preserved by the *Conservatism* of the revision. Table 1 shows two common examples: Bayesian updating and Jeffrey updating. Note that the propagation step covers what is typically thought of as a "belief revision rule," such as updating by Bayesian conditioning.

Table 1: Two belief revision rules

Rule	Perturbation	Propagation
Bayes	$Q(E) = 1$	$Q(X) = P(X\|E)$
Jeffrey	$Q(A) = \pi_A, \forall A \in \mathbb{A}$	$Q(X) = \sum_{A \in \mathbb{A}} P(X\|A)Q(A)$

Our first pass followed this pattern. Perturbation: the report sets a constraint, $Q(R) = 0.6$. Propagation: Jeffrey condition on the parti-

[12] I take this term from Bradley (2017).

tion $\{R, \neg R\}$ to restore coherence. But, clearly, expert deference isn't a *kind* of belief revision, as Bayes and Jeffrey updating are, for in example 2 it worked as an *instance* of Jeffrey updating. This is why I call my proposal expert deference as a belief revision schema.

In generalised belief revision theory, a *kind of experience* is matched with a *particular revision rule*. Dietrich, List, and Bradley (2016) characterise the class of Bayesian inputs as those experiences which constrain the agent's belief state to include only probability functions in which the probability of a specific proposition—the one the agent learns during the experience—is 1. We can similarly define the class of Jeffrey inputs (corresponding to Jeffrey updates) or Adams inputs (corresponding to Bradley's Adams updates (Bradley, 2017)). In each case, we can model this with a domain $\mathcal{D}$ that is a subset of the space of possible experiences and initial states: $\mathcal{D} \subseteq \mathcal{P} \times \mathcal{I}$.

Responsiveness consists in ensuring that the final belief state is in the set I, i.e., that it is compatible with the experience. Conservatism is harder to spell out. Each domain that Dietrich, List, and Bradley consider comes with a specification of what those experiences are "silent" on. This notion of silence is used to fill out the norm of Conservatism: put roughly, a belief revision rule should leave unchanged whatever the experience is silent on. Dietrich, List, and Bradley (2016) then prove characterisation results showing that, for each of these three domains, there is a unique rule respecting Responsiveness and Conservatism and that in each case it is the rule referred to parenthetically above.

But experts may plausibly report many kinds of probabilistic information. They might tell you that one event is more likely than another, or that two variables are independent, or they might specify the expected value for some variable. (van Fraassen (1981) says that reporting the expected value for a variable is the most general kind of constraint on your probability function, and that others can all be framed as special instances of it.) We want our theory of expert deference to be able to handle all of these report types.

The problem is that for these more general input domains, no unique belief revision rule is known. Put extensionally, the problem is that once we identify the set of belief states that respect the input, we lack general rules for further refining this set. This may be easier to see by switching to an intensional definition. Let us denote the constraint imposed by the expert report with a formula, ϕ_Q. Responsiveness tells us that the posterior credence function must respect this constraint: we want a Q that respects ϕ_Q. But there are a great many of these! What we want is one which also fits the prior, P, in the right way. What is that way? Conservatism is meant to provide the answer: in the way that preserves as much of P as possible while respecting Q.

In practice, specifying the Conservatism norm for a form of experience is a complex matter. The canonical examples mentioned above have a particularly nice form: each comes with a constraint and a *rigidity* condition which realises the Conservation condition. These conditions, summarised in Table 2, are necessary and sufficient conditions for updating according to the associated belief revision rule

(Bradley, 2017, pp. 188-200). (Adams updating will be introduced below.)

Table 2: Conservatism conditions for Bayes, Jeffrey and Adams updates. $\mathbb{A}, \mathbb{B}$ are partitions of propositions, and the π's are particular probability values.

	Constraint	Rigidity condition(s)
Bayes	$Q(E) = 1$	$Q(\cdot\|X) = P(\cdot\|X), \forall X \in \Omega$
Jeffrey	$Q(A) = \pi_A, \forall A \in \mathbb{A}$	$Q(\cdot\|A) = P(\cdot\|A), \forall A \in \mathbb{A}$
Adams	$Q(A\|B) = \pi_A^B,$	$Q(\cdot\|AB) = P(\cdot\|AB), \forall A, B \in \mathbb{A}, \mathbb{B}$
	$\forall A, B \in \mathbb{A}, \mathbb{B}$	$Q(A) = P(A), \forall A \in \mathbb{A}$

In the general cases discussed above, we don't have Conservation conditions that produce unique "kinematic" update rules of this kind. We either need to do more work to identify what is conserved by the experience (or what it is "silent" on, in the language of Dietrich, List, and Bradley, 2016), in order to formulate a kinematic revision rule, or, lacking that, we need an alternative way of getting from P to the right kind of Q. In section 2.7 I will discuss one approach, divergence minimisation, that I do not think is fruitful.

2.4.3 *Three kinds of deference*

This way of thinking of deference is flexible enough to handle a variety of kinds of reports, while retaining the benefits of familiar belief revision rules. This is because it is a schema that picks out these familiar rules for particular kinds of report. The simplest case is one where an expert tells you that some proposition is true. For example, perhaps a completely reliable person looks outside and tells you it is raining. You defer to them, and set your credence in this proposition to 1. This is an instance of expert deference that provides a Bayes-type constraint on your credences and so the schema tells you to update your beliefs using the appropriate rule: Bayesian conditioning. You haven't undergone a Bayesian learning event and so this isn't just Bayesian learning—after all, you haven't observed the rain, and the proposition you learned is $\ulcorner R \urcorner$ rather than R itself.

Let us look in some more detail at three more interesting kinds of expert reports, for which our schema will recommend the use of Jeffrey conditioning, Adams conditioning and odds kinematics—each of which will be more fully introduced below.

Unconditional Deference

We start with the familiar kind of reports, discussed in section 2.3. As mentioned there, deference is often taken to be this simple principle, relating to the unconditional probabilities of propositions.

Definition. (Unconditional Deference) Consider some partition $\mathbb{X}$ in the expert domain Ω. If an expert reports $P^i(X) = x^i$, for each $X \in \mathbb{X}$ then the agent should set their probabilities to $Q(X) = x^i$.

Note the reformulation of the principle: there is no discussion of priors, and the principle is framed as a normative imperative for the agent. It partially specifies a belief revision: an agent who follows this principle will change those particular probabilities $P(X)$ to $Q(X)$ for all $X \in \mathbb{X}$. This provides the perturbation for a belief revision. What of the Conservatism condition, and associated propagation? A suggestion advanced by Katie Steele is that we should perform a Jeffrey update on the partition $\mathbb{X}$ (Steele, 2012, pp. 990-92).

Steele's suggestion is a natural one, as Jeffrey conditioning is a widely accepted rule for restoring coherence upon receiving new unconditional probabilities over a partition from a Jeffrey experience.[13] In an Unconditional Deference situation, we also begin with new probabilities over a partition. Further, what the novice learns from the expert is assumed to be *entirely* represented by the set of new probabilities $\{P^i(X)\}_{X\in\mathbb{X}}$. Our expert was (literally) silent on other matters, including probabilities conditional on the X's. So we can safely assume that the Rigidity condition, which is the Conservation condition for Jeffrey experiences, holds.

So we can complete our belief revision by having the novice first set $Q(X) = P^i(X)$ for all $X \in \mathbb{X}$ by Unconditional Deference, and then adjust the rest of their beliefs by Jeffrey conditioning on the partition $\mathbb{X}$, using Eq.1. (This "first. . . then. . . " language is somewhat metaphorical. We can think of the two steps as being separated logically, rather than representing an actual sequence of changes to the agent's partial beliefs.)

As mentioned above, Jeffrey conditioning is not without its opponents, particularly because it is often said to be *non-commutative*—i.e., when considering two Jeffrey updates, we get different final probabilities depending on the order of the updates. This is not actually correct for Jeffrey *experiences*. The non-commutativity arises from simple applications of Jeffrey's rule, without paying proper attention to the

[13] The rule has been given a Dutch Book justification by Skyrms, and is popular. This is not to say it is without controversy, and I discuss some of the concerns regarding it below.

experience generating the changes. Let us say an agent “updates on a partition $\mathbb{X}$” to mean that they propagate probabilities for that partition across their probability function using Jeffrey’s rule, denoted $P_{\mathbb{X}}$. If we describe an agent updating first on a partition $\mathbb{X}$, then on another $\mathbb{Y}$, the result will in general differ from that derived from first updating on $\mathbb{Y}$ and then $\mathbb{X}$: $P_{\mathbb{X}\mathbb{Y}} \neq P_{\mathbb{Y}\mathbb{X}}$. (The easiest way to see this is to consider two updates on $\mathbb{X}$, one with the profile of probabilities (x_j^1) and the other with (x_j^2). The probabilities for $\mathbb{X}$ will end up as those from the *second* update, which simply replace the first.) This is a fact of the mathematics of Jeffrey conditioning. Nevertheless it isn’t right to say the result of *sequential Jeffrey experiences* doesn’t commute, because this is not the proper representation of sequential Jeffrey experiences. A careful analysis of sequential Jeffrey experiences, such as that provided by Wagner (2002), allows us to model them in a *commutative* manner.[14]

However, while successive Jeffrey *experiences*, properly modelled, will have *commutative* effects, the procedure Steele recommends for updating on testimony may be *non-commutative*.[15] This is because it *cannot* be understood in terms of sequential Jeffrey experiences. There is no ineffable learning experience going on here; all we are doing is

[14] Wagner presents his results as an extension of those which stipulate sufficient conditions for Jeffrey conditioning commuting. For Wagner, these conditions are two identities for the Bayes factors generated by each of two experiences, relative to two different partitions $\mathbb{E}$ and $\mathbb{F}$. Let $\beta_{1\mathbb{E}}$ represent having the 1st Jeffrey experience occurring relative to partition $\mathbb{E}$. Then Wagner’s identities are $\beta_{1\mathbb{E}}(E_i, E_j) = \beta_{2\mathbb{E}}(E_i, E_j) \forall i, j$ and the same for $\mathbb{F}$. But as Wagner notes, Bayes factors are the right way of representing what is learned in an experience in a prior-free way. So if we stipulate that the experiences are identical, then Wagner’s Bayes factor identities hold, and therefore Jeffrey conditioning commutes across the order of the experiences.

making use of the machinery of Jeffrey conditioning in this particular instance of the belief revision schema of expert deference.

This should not surprise us in the context of Expert Disagreement. The general lack of commutativity is precisely the problem of expert disagreement, realised in our model: the mechanism I have outlined thus far for integrating an expert report—unconditional deference followed by Jeffrey conditioning—has made no advances on the source of the problem of Expert Disagreement. Unconditional Deference treats all experts on the same level, as though they were the same expert. In the next section I will provide a fix that works for all three forms of deference discussed here.

Conditional Deference

Conditional probabilities capture the relevance relations between propositions, a vital constituent of expertise. They are therefore an extremely common part of expert elicitation. More prosaically, medical prognoses often involve doctors telling us the likelihood of recovery given a diagnosis, and often the diagnosis is itself (implicitly) prob-

[15] Why only "may be" non-commutative? In our finite setting, Diaconis and Zabell (1982) provide necessary and sufficient conditions for commutativity. Consider two partitions and the sequences of probabilities assigned to them in a Jeffrey update: $\{\mathbb{X}, \langle x_j \rangle\}$ and $\{\mathbb{Y}, \langle y_k \rangle\}$. $\mathbb{X}$ and $\mathbb{Y}$ are *Jeffrey independent* with respect to P, $\langle x_j \rangle$ and $\langle y_k \rangle$, if $P_{\mathbb{X}}(Y_k) = P(Y_k)$ and $P_{\mathbb{Y}}(X_j) = P(X_j)$ holds for all j, k. Then successive Jeffrey updates commute, $P_{\mathbb{X}\mathbb{Y}} = P_{\mathbb{Y}\mathbb{X}}$, if and only if $\mathbb{X}$ and $\mathbb{Y}$ are Jeffrey independent with respect to P, $\langle x_j \rangle$ and $\langle y_k \rangle$ (Diaconis and Zabell, 1982, Theorem 3.2). This turns out to be a weaker condition than probabilistic independence, so that if $\mathbb{X}, \mathbb{Y}$ are probabilistically independent with respect to P, successive updates on them will commute for any update probabilities (Diaconis and Zabell, 1982, Theorem 3.3). (Jeffrey independence can also be stated in terms of Bayes factors, which is an essential part of Wagner's argument for the commutativity of sequential Jeffrey experiences—see footnote 14.) So, while some sequences of Jeffrey updates will commute, in general we should expect them not to.

abilistic and conditional on various symptoms and test results. We typically, and I say rationally, defer to these conditional opinions.

Definition. (Conditional Deference) Suppose that an expert reports their conditional probability for X given some other possibility Y: $P^i(X|Y) = x^i_y$, for $X, Y \in \Omega$. Then the agent should set their conditional probability to $Q(X|Y) = x^i_y$.

Once again, this provides us with a constraint that can act as a perturbation for a belief revision rule. I will work with the following medical example in fleshing out exactly what that rule should be.

Example 3. *Eva has a family history of ovarian cancer. She has been told that, based on this, she has a 3-5% lifetime chance of developing the cancer herself. Recently, she learned that there is a genetic mutation called BRCA2, present in* $1/1000$ *people, which makes ovarian cancer more likely. She meets with a genetic counsellor, who tests her for the mutated gene. Before she receives her results, the doctor tells her: "patients with a mutated BRCA2 gene have a much higher risk of developing ovarian cancer, around 23%." If she tests negative, her risk will be in the range she was previously told, 3-5%.*

Let O be the proposition that Eva will develop ovarian cancer at some stage in her life, and B stand for having a mutation on that gene. Assume that coming in, she has $P(O) = 0.04$ and $P(B) = 0.001$. Given the doctor's testimony, it seems reasonable for Eva to set $Q(O|B) = 0.23$ and $Q(O|\neg B) = 0.04$. Her unconditional probability $P(O)$ should change, though it isn't immediately clear how. What does seem clear is that she *shouldn't* adjust the probability that she has the mutated gene $P(B)$—she's received no information relevant to this (yet; she will receive her test results in time).

What Eva needs is a way of updating her beliefs that respects these constraints. This will involve adjusting her joint probabilities across the possibilities: developing cancer, not developing cancer, having the mutated gene, not having it. Realistically, Eva didn't have any views at this fine grain before (it seems implausible that she had a conditional probability of developing cancer given a positive test result), so it is more accurate to say that what she needs to do is to distribute her beliefs over the possibilities of developing ovarian cancer and having the mutated gene in a way that fits the expert testimony she has received.

Richard Bradley (2005) described a procedure called *Adams conditioning* for changing the conditional probability for X given some possibility Y, without changing the probability for the possibility Y itself. In fact, as with Jeffrey conditioning, Bradley's procedure is defined in terms of partitions.

Definition. (Adams conditioning) Suppose that an agent considers two partitions, $\mathbb{X}$, $\mathbb{Y}$, and that they come to have new conditional probabilities for the elements of $\mathbb{X}$, given the elements of $\mathbb{Y}$, denoted π_j^k for each $X_j \in \mathbb{X}$ and $Y_j \in \mathbb{Y}$. Their new probability function Q is obtained from P by Adams conditioning if and only if, for all $Z \in \Omega$:

$$Q(Z) = \sum_k \left[\sum_j P(Z|X_j Y_k) \pi_j^k \right] P(Y_k) \tag{2}$$

Adams conditioning is the right way to update when two conditions hold. The first we have already seen: $Q(Y_k) = P(Y_k)$ for each k, which Bradley calls *Independence*. The second is another Rigidity condition, this time for the probabilities conditional on the cells of the joint partition $\mathbb{XY}$: $Q(\cdot|X_jY_k) = P(\cdot|X_jY_k)$. These jointly constitute the Conservation condition for this belief revision rule. The repetition of the familiar Rigidity condition provides some of the motivation for this revision procedure: Adams conditioning can be thought of as a special case of Jeffrey conditioning. As Bradley puts it, "if Jeffrey conditioning is the correct revision rule for Jeffrey experiences then Adams conditioning is the correct rule for Adams experiences" (Bradley, 2017, p. 178).

In Eva's case, this means that her probability for a general proposition Z should now be:

$$\begin{aligned} Q(Z) = {} & [P(Z|OB)Q(O|B) + P(Z|\neg OB)Q(\neg O|B)]\, P(B) \\ & + [P(Z|O\neg B)Q(O|\neg B) + P(Z|\neg O\neg B)Q(\neg O|\neg B)]\, P(\neg B) \end{aligned}$$

If we consider $Z = O$ and ask what her probability in developing ovarian cancer should be, we get:

$$Q(O) = Q(O|B)P(B) + Q(O|\neg B)P(\neg B)$$

This is just the law of total probability, but with $Q(B) = P(B)$ due to the Independence condition. This is the natural way to set $Q(O)$ given what Eva has available to her at this stage in the example. Given the tiny probability of having the mutated gene $P(B)$, we can see that Eva should have $Q(O) \approx Q(O|\neg B) = 0.04$.

As Adams conditioning can be thought of as a special case of Jeffrey conditioning, it is no surprise that in general its sequential application here will be non-commutative. The reflections at the end of the discussion of Unconditional Deference above therefore also apply here.

Experiential Deference

The two forms of deference above concern expert's sharing specific probabilities, and an agent incorporating those into their own probabilities. However, there is at least one other form of deference to expertise discussed in the literature on Bayesian models of expertise. Richard Jeffrey introduces the final kind that I will consider by considering cases in which an agent wishes to use a probabilistic report without simply adopting it as their own, as such reports are "necessarily a confusion of what the other person has gathered from the observation itself, which you would like to adopt as your own, with that person's prior judgmental state, for which you may prefer to substitute your own" (Jeffrey, 2004, p. 59).

For our purposes, this mode of deference, which I will call *experiential*, is of particular value when considering experts consulting one another, and wishing to defer in a selective manner.

Jeffrey was motivated to consider such cases by his work on pathology, in which expertise is highly differentiated. A pathologist might be an expert in arriving at diagnoses (elements of $\mathbb{D}$) by performing tests, but a physician may be an expert on the prognoses resulting from these diagnoses (represented as conditional probabilities for $\mathbb{H}$ given $\mathbb{D}$). Now in some cases the physician may wish to simply adopt the probabilities provided by the pathologist, which they can do by Unconditional Deference. This can happen even in the absence of any unconditional priors of their own on $\mathbb{D}$—which Jeffrey takes to be realistic given the specialisation in medicine—so long as they have their own conditional priors $P(X|D)$ for all $D \in \mathbb{D}$ and $X \in \Omega$.

But, says Jeffrey, a clinician may also have their own priors over the D's, which they wish to use in forming a judgement. (Perhaps these incorporate information about this patient that the pathologist does not have.) It would therefore be inadvisable to use Unconditional Deference, which would "contaminate" this judgement with the pathologist's own priors over $\mathbb{D}$. Stepping outside of the language of the Bayesian model for a moment, what we need is a way to convey the impact of the *experience itself* from pathologist to physician. The pathologist will ideally communicate how one's views on the diagnosis ought to change, given the evidence in the test, *no matter what those views are*.

Jeffrey proposes the following procedure. The pathologist's new and old probabilities over $\mathbb{D}$ can be used to calculate their Bayes factors over the partition, compared to some fixed element which we call D_1:

$$\beta(D_j, D_1) = \frac{Q(D_j)/Q(D_1)}{P(D_j)/P(D_1)}$$

Since at least 1950, Bayes factors have been recognised as the right tool for representing what is learned during an experience in a *prior-*

free way (Good, 1950). These, then, are the right things to defer to in a case of Experiential Deference.

Experiential Deference differs from Unconditional and Conditional Deference in that it is defined relative to an instance of *learning* on the part of the expert. The expert here reports on a learning experience, rather than simply sharing their view on a proposition of interest. Learning is a crucial question for formal models of expert disagreement, as we will see again when we consider opinion pooling axioms like External Bayesianity in chapter 3.

Jeffrey developed a rule for updating one's probabilities using Bayes factors (Jeffrey, 1992; Jeffrey and Hendrickson, 1989) that he called *odds kinematics*.

Definition. (Odds kinematics) An agent can use expert i's Bayes factors (β^i) to update their probability, from $P \mapsto Q$ in the following way. For an arbitrary proposition Y, the agent sets the odds on that proposition to

$$\frac{Q(Y)}{Q(\neg Y)} = \frac{\sum_k \beta^i(X_k, X_1)P(Y|X_k)P(X_k)}{\sum_l \beta^i(X_l, X_1)P(\neg Y|X_l)P(X_l)} \tag{3}$$

An agent who has good reason to preserve their own priors may defer to an expert's skill at judging experiences using odds kinematics.

2.4.4 *Summary so far*

I have introduced a novel way of thinking of expert deference: as a belief revision schema. That is, we model expert testimony as providing an exogenous constraint on an agent's posterior credences, and use a belief revision rule that matches the form of this constraint to complete an update.

My model removes the expert reports from the algebra Ω entirely. It also thereby removes any priors over propositions like $\ulcorner W \urcorner$. I have also removed the role of the agent's priors for the proposition the report *concerns*, e.g., H in Ex. 1, or R in Ex. 2. In this way, the new model significantly reduces the Cognitive Burden placed on rational agents, and resolves the problem of the Relevance of Priors to expert domains.

The new proposal, as it is an instance of expert deference, is Sensitive to Testimony. It uses the content of the report *directly*, recognising the kind of experience expert testimony is. Because it is a belief revision *schema*, it avoids the problem of Arbitrariness, as it places no special focus on unconditional probabilities. Similarly, it inherits the "good behaviour" of beloved belief revision rules like Bayes, Jeffrey and Adams updating in the appropriate cases.

In hard cases, it faces a number of challenges. There is no general solution to the problem of belief revision. But this is not a problem particular to my model, as it has nothing to do with the particularities of expert testimony. The problem emerges because the reports are only partial constraints on credences. The result of my procedure may therefore be that agents end up with quite imprecise posteriors, as we shall see in the following section. But this should not be taken as failure! If the information received from an expert does not provide very tight constraints on our credences, and we respect those reports, we should expect to end in an imprecise state.

A final note on imprecise probabilities: I have considered different kinds of probabilistic report, in the sense of the expert reporting unconditional probabilities, conditional probabilities, or Bayes factors. But I assumed that these reports were *precise* in all cases: the expert's reported single numbers. However, the model could easily be extended to cover imprecise probabilistic reports, of any kind. The belief revision framework that I use involves determining the "input" that the learning experience provides. There is no reason to suppose that this input must be precise. The expert report could just as easily require that the agent's posterior probability for the relevant proposition lie within a range. More would need to be said about how to conduct the propagation step for such an imprecise input, but there do not seem to be new difficulties here beyond those already found in the imprecise probabilities literature. So, if one is willing to accept imprecise probabilities at all, my approach can be adapted to handle them.

2.5 awareness growth

Awareness and Expert Disagreement are more difficult issues, and each has unique challenges. In this section I will show how expert deference as a belief revision schema can go together with one model for rational awareness growth, and in the next section I will address disagreement.

Let us return to Ex. 1: *You open your weather app and see, to your complete surprise, that there is a 30% chance that London will be struck by a hurricane on Thursday.*

In reviewing this example above, I said that one implausibility of supra-Bayesianism is that the agent had never considered this possibility before. They simply have no attitude toward it. This will be my definition of unawareness: a state in which an agent has no attitude to a proposition. Therefore, in our model, it is natural to represent this with a proposition, e.g., H, that is not in the domain of P. As P's domain is the whole algebra Ω, this means that $H \notin \Omega$

We know where we would like to end up: the agent comes to have an attitude to H, in particular $Q(H) = 0.3$. The problem of awareness growth is to find rational constraints on Q, linking it with the prior belief state P.

Here is how my proposal goes. We can think of the expert's report as having two parts: the proposition it concerns (which I will sometimes call its *content*), in this case H, and the probabilistic information it conveys about that content (which I will call its *value*), in this case 30%. I will separate my treatment of awareness growth into two stages, corresponding to these two parts. (The separation is conceptual, and should not be taken to imply that the agent follows this sequence.) The first stage is purely a matter of awareness: the agent becomes aware of the new proposition(s) and we determine how their old attitudes can be extended to an algebra that contains the new proposition(s). The second stage is a matter of learning: the agent became aware of the proposition(s) via a learning experience, and in particular an experience of expert testimony. This gives them evidence about the new propositions. In the second stage, I will apply my expert deference proposal, show how it fits naturally with one theory of rational awareness growth.

I will start with some discussion of awareness growth in general, to set the stage. Bradley (2017, pp. 256–8) differentiates between two kinds of awareness growth, distinguished by the kind of belief change they require of the agent.[16]

First, an agent may come to realise that the possibilities they considered were too coarse. *Refinement* involves making new distinctions, dividing up the possibilities into finer units. Suppose I am considering the weather, and I initially entertain just two options: rain (R) or shine (S). I later realise that temperature is important too. I distinguish two temperatures, warm (W) and cold (C), and recognise that it can be rainy and warm, rainy and cold, sunny and warm, and sunny and cold. (At this point in the story, I know nothing about the relations between temperature and precipitation.) Note that because this is a refinement, $\{R, S\}$ is still a partition for me ($R \vee S = \top$). What has changed is that R and S are no longer "primitive possibilities," in the sense that they aren't maximally specific: I am now aware of two ways that it might rain, RW and RC. The primitive possibilities are now the four elements of the joint partition $\{RW, RC, SW, SC\}$.

[16] Bradley takes these two options to be exhaustive; other forms of awareness growth are reducible to combinations of them. I do not need this to be true in what follows.

In the second kind of awareness growth, *expansion,* the agent becomes aware of a new primitive possibility. Suppose again that I am considering the weather, and I initially entertain just two options: rain (R) or shine (S). Then, upon looking at my app, I consider a third option which I see as mutually exclusive of these two: hurricane (H). Before, I regarded $\{R, S\}$ as a partition: $S = \neg R$ and $R \vee S = \top$. After my awareness grows, $R \vee S$ is merely a contingent proposition, $\neg R = S \vee H$, and $\{R, S, H\}$ is a partition.

Here is how I propose to model these learning experiences. First, we deal with the purely awareness related part of the experience. We will "grow" the algebra Ω to include the content. Then we will extend the agent's prior credence function P to the new, wider algebra $\Omega^{\oplus}$. The thought here is that some aspects of P will be preserved in a process of rational awareness growth, and so we can specify some conditions that any credence function on $\Omega^{\oplus}$ must have in order to "fit" with P. Second, we model the learning experience. Here I bring in my belief revision schema, beginning with a perturbation using the report: the probability for H is set by the report's constraint ϕ_Q. Finally, I "fill in" this Q to "match" the extended prior, in a move analogous to the propagation part of a belief revision process. (All of the terms in scare quotes are loose descriptions that will be made more precise as we go.) For clarity I will separate each stage—awareness change and learning—into two steps.

Step 1: Growing the algebra.

Having outlined what awareness growth is, we can consider how to revise belief in the fact of expert testimony on novel propositions. The first step makes use of only the content of the testimony, setting aside its value. We start by forming a new algebra, containing all the propositions the agent was previously aware of and the propositions reported on.

Making this more precise requires slightly more mathematical machinery than we have thus far employed (following Bradley, 2017, pp. 258–9). I have been using a logical framework in which Boolean algebras are lattices of propositions, ordered by an implication relation. In order to make the lattice structure explicit, I will now write $\Omega = \langle \mathcal{X}, \models \rangle$, where $\mathcal{X}$ is a set of propositions and $\models$ is the implication relation that acts as the order for the lattice. The top element of the algebra is typically denoted $\top$, but in the context of multiple algebras it is useful to think of it as merely being the upper bound of the set $\mathcal{X}$: $\vee\mathcal{X}$.

In general we can suppose that the agent becomes aware of a set of propositions $\mathcal{U}$, with $U \notin \mathcal{X}$, for all $U \in \mathcal{U}$. We start by forming Y, the closure of $\mathcal{U} \cup \mathcal{X}$ under the Boolean operations. Then $\Omega^{\oplus}_{\mathcal{U}} = \langle \mathcal{Y}, \models \rangle$ is a Boolean algebra, which Bradley calls the extension of Ω by $\mathcal{U}$. Note that $\vee\mathcal{X} \in \mathcal{Y}$, and in general $\vee\mathcal{X} \neq \vee\mathcal{Y}$.[17]

[17] Note a persistent idealisation here: $\models$ is the implication relation which ordered the old algebra, and it also orders the new propositions. So, the agents that we model in this framework are logically omniscient (as is standard) and this omniscience extends to propositions they were previously unaware of. The problem of logical omniscience is a significant one for someone with my non-ideal theory interests. However, treating it is notoriously difficult. I therefore put up with this idealisation,

The old algebra is related to the new one via an embedding. A lattice embedding is a one-to-one homomorphism: a function that maps each proposition in the old algebra to a proposition in the new algebra, and which preserves the lattice operations, meet and join—which is to say, logical conjunction and disjunction. It does *not* preserve logical complements.

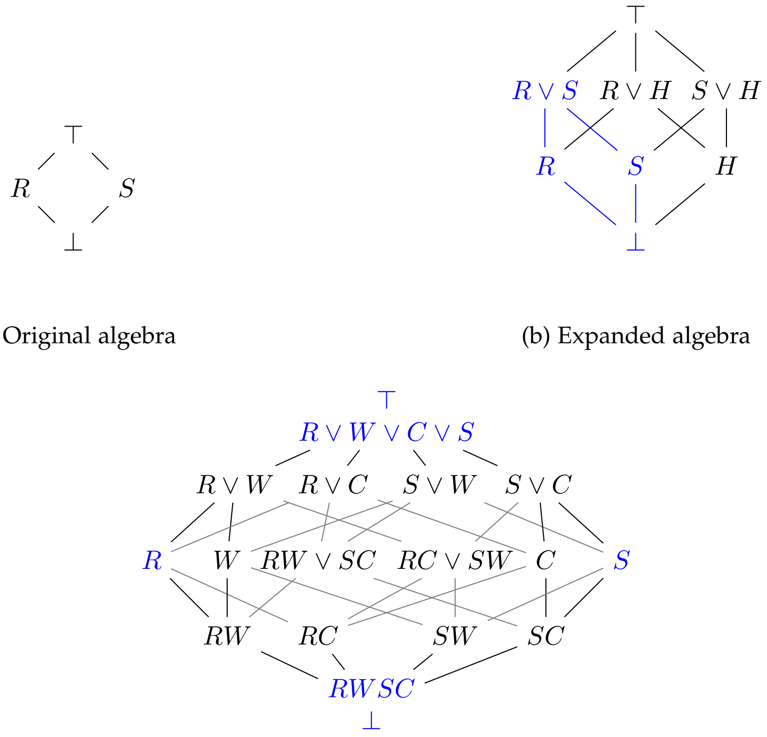
We can see this by considering the weather example again. Simple lattices can be usefully visualised with Hasse diagrams, such as those in figure 1. The lines connect the higher-up propositions with the logically stronger propositions beneath them which entail them. The figure highlights in blue the elements of the new algebra which correspond to the old propositions. Note that the element $R \vee S$ is in $\Omega^{\oplus}$ in both the expansion and refinement case. In the expansion case it is now merely a contingent proposition, as the tautology in $\Omega^{\oplus}$ is $R \vee S \vee H$. In figure 1(b) it is easy to see that the old algebra is a sub-algebra of the new. In the refinement case, $R \vee S$ is also (or, if you prefer, still) the top element of $\Omega^{\oplus}$. Figure 1(c) shows how much more complicated things look with more propositions.

Step 2: Extending the probability function.

What we have now is an algebra, $\Omega^{\oplus}$, which contains the new propositions. But P is not defined on $\Omega^{\oplus}$, but rather on Ω. Our second task is to extend P to $\Omega^{\oplus}$. "Extension" is not belief revision. I am interested here merely in what the prior P has to say about $\Omega^{\oplus}$.

noting that allowing agents to be unaware does mitigate force of the problem of logical omniscience.

Figure 1: Hasse diagrams showing Bradley's two kinds of awareness growth.



(a) Original algebra

(b) Expanded algebra

(c) Refined algebra

There are different proposals for how to extend a probability function to a wider algebra. Two prominent proposals are "Reverse Bayesianism" (due to Karni and Vierø, 2013) and "Rigid Extension" (due to Bradley, 2017). I will employ the latter method.

Bradley (2017, p. 257) begins by considering simple examples of refinement and expansion, like my weather case. He characterises refinement as coming to realise that the possibilities I previously considered are too coarse-grained. But introducing finer-grained possibilities should not change my attitude to the coarse-grained possibilities;

realising that there are two kinds of rainy weather should not change the probability of rain overall.

For expansion this is not the case. Introducing an entirely new kind of weather must change my attitude to at least one of the possibilities I previously considered, since my degrees of belief must sum to one. But, Bradley argues, there should not be any relative change between the old propositions: if I previously thought rain and shine equally likely, I have no reason to alter that relative comparison now that I have discovered that those alternatives do not exhaust the possibilities.

The core idea here is simple, and in line with our discussion above: minimal change. Bradley argues that the Conservation condition for three Bayesian belief revision rules (Bayes, Jeffrey and Adams updating) involves the rigidity of conditional beliefs. So, he concludes, a Conservation condition for extension to a wider algebra should also preserve conditional probabilities. He provides such a condition which captures the intuitions that he defends about refinement and expansion cases.

Specifically, "the agent's new conditional probabilities, given the old domain, for any members of the old domain should equal her old unconditional probabilities for these members." (Bradley, 2017, p. 258) Or, to use terminology Bradley introduces, the new belief states must be rigid extensions of the old.

Definition. (Rigid Extension) A probability function $P^{\oplus}$ on $\Omega^{\oplus}$ is called a *rigid extension* of P iff, for all $X \in \Omega$, $P^{\oplus}(X | \bigvee \mathcal{X}) = P(X)$.

In general, there will be many rigid extensions of a credence P to a wider algebra $\Omega^{\oplus}$. Rigid Extension concerns only the parts of the new function that involve propositions from the old algebra, so it leaves open many possible assignments of probability to the new propositions, and combinations of new and old. (Rigid Extension does constrain the latter.) In figure 1, these elements of the old algebra are shown in blue, when embedded in the new algebras on the right.

This is a minimal condition, and will therefore typically result in imprecise posterior credences even if one starts with a precise credence P on Ω. I will denote the result of Bradley's procedure $\mathbb{P}^{\oplus}$; it is the set containing all the rigid extensions of P to $\Omega^{\oplus}$.

Step 3: Perturbation.

$\mathbb{P}^{\oplus}$ is an intermediate construct. It represents what the agent's old probabilities have to say about the new possibilities the agent is aware of. We can now model the learning experience that the agent undergoes in virtue of *hearing an expert report* about those new possibilities.

As before, the report provides a constraint ϕ_Q on any credence function that is Responsive to the learning experience. Note that in this case a single experience is producing both the growth of the algebra and the perturbation of the credence function. The decomposition into two steps is merely a logical decomposition, helpful in modelling awareness growth—there is no implication that the two are separated in time, or occur separately.[18]

[18] Bradley's approach fits naturally with my topic of expert deference. I want the posterior attitudes to the new propositions to come from the expert reports, and

The constraint in example 1 is that $Q(H) = 0.3$, and so we will work with the set of potential functions $\mathbb{Q} = \{Q$ is a credence on $\Omega^{\oplus}$: $Q(H) = 0.3\}$.

Step 4: Propagation.

We now want to further constrain $\mathbb{Q}$, so that it fits with our extended prior $\mathbb{P}^{\oplus}$. In our simple belief revision cases above, we accomplished this by using a Conservation condition that told us which parts of the prior should be conserved. We used these conserved quantities, together with ϕ_Q, to fix the posterior (uniquely, in cases of kinematic belief revision).

As discussed above, the Conservation condition depends on the nature of the learning experience. In our example, the report provides us with an unconditional probability for H. When we encountered reports of this kind before, we were able to motivate for the rigidity of conditional probabilities as the Conservation condition, and therefore to use Jeffrey conditioning as our propagation procedure. We *cannot* do that here. The agent was unaware of H before the learning experience, and so does not have prior conditional probabilities concerning H.[19] As they don't exist, they cannot be the subject of any Conservation condition such as rigidity. A similar argument blocks

therefore it is helpful to maintain the separation between these steps. (Karni and Vierø, 2013) do not provide anything like this clean separation. In other cases, this methodological separation may be not be desirable: in an unpublished manuscript Stefánsson and Steele (ms) argue that this two-step procedure is baseless. In my case I think the value of the separation is clear.

[19] Put another way: $\mathbb{P}^{\oplus}$ has no constraints on conditional probabilties involving H that aren't just consequences of Rigid Extension.

the use of any kinematic rule for any awareness-growing instance of expert testimony. The challenges we face here are therefore akin to those discussed in section 2.4.2 for general reports.

There will be some cases where we can make progress. If the algebra is simple enough, and the constraint ϕ_Q restrictive enough, we might nevertheless determine a unique posterior Q. I present two simple cases, to illustrate the difficulties and how they can sometimes be avoided.

Example 4. *Expansion: Considering the weather, you initially entertain the possibilities* $\{R, S\}$*, assigning each credence* 0.5*. You later become aware of* H*, which you take to be a distinct possibility, when you hear the weather report,* $W(H) = 0.3$*.*

Let's work through the four steps.

1. *Grow algebra:* Your initial algebra Ω is generated by the two-element partition $\mathbb{A} = \{R, S\}$. The expanded algebra $\Omega^{\oplus}$ is generated by the three-element partition $\{R, S, H\}$. It has the structure shown in 1(b).

2. *Extend prior:* Your prior P is fully specified by $P(R) = 0.5$. A function $P^{\oplus}$ on $\Omega^{\oplus}$ is a rigid extension of P iff $P^{\oplus}(X| \vee \mathcal{A}) = P^{\oplus}(X|R \vee S) = P(X)$. So, your extended prior consists of all such functions, for which $P^{\oplus}(R|R \vee S) = 0.5 = P^{\oplus}(S|R \vee S)$.

3. *Perturb:* Here we simply form the set of all Q's on $\Omega^{\oplus}$ such that $Q(H) = 0.3$.

4. *Propagate:* To make progress we consider the structure of the algebra. This testimony has introduced a third primitive propo-

sition, making $\{R, S, H\}$ a partition. The constraint $Q(H) = 0.3$ determines the probability of $Q(R \vee S) = 1 - Q(H) = 0.7$. By rigid extension we know $Q(R|R \vee S) = 0.5$, and this is sufficient to fix a unique Q, specified by the following probabilities for the atoms: $Q(R) = 0.35 = Q(S), Q(H) = 0.3$.

We are able to make progress here because between the report and the demands of rigid extension, we fix the probabilities of the atoms of the new algebra. This will not always be the case.

Example 5. *Refinement: Considering the weather, you initially entertain the possibilities* $\{R, S\}$*, assigning each credence* 0.5*. You later become aware of temperature,* $\mathcal{E} = \{W, C\}$*, which you know can combine with R, S, when you hear the weather report* $W(W) = 0.6$*.*

Things are harder in the refinement case. We can follow the steps just described, but there is an important difference. Refinement changes the atoms of the algebra: that's part of what it means to fine-grain in this setup. The atoms of the new algebra are the elements of the finest joint partition over all currently known partitions. (These are the maximally specific possibilities.)

If ϕ_Q is specified at the level of the new algebra's atomic propositions —in the example, $\{WR, WS, CR, CS\}$—and determines the probabilities for all atoms, then Q will be uniquely specified. If not, it will be under-specified. There are many ways that a probability function can satisfy $Q(W) = 0.6$: temperature might be independent of precipitation, or warmth may be more likely if there is no rain, or warmth may be more likely if there *is* rain.

If the constraint is at the level of the new coarse-grained partition, rather than the joint partition, the result will be an imprecise posterior $\mathbb{Q}$, containing all the joint distributions meeting the constraint.

We can now see how my proposal for expert deference as a belief revision schema fits naturally with this approach to awareness growth. The fact that the agent is unaware of the content of the expert report means that more orthodox Bayesian methods could not make progress here. It is implausible that the agent has a prior for the reported proposition, or any attitudes about the proposition that the expert make the reports they do about the content. In my proposal, this is no problem. Expert reports are not within the model, but rather act as external constraints on posterior credences. It does not matter that the propositions these reports concern are not in the old algebra, as they only play a role after we have formed a new algebra and extended the agents priors to that new algebra.

My discussion highlights the difficulties of rational awareness growth. When one becomes aware of a new proposition, one needs to consider not only one's attitude to that proposition but also how it relates to all the familiar propositions. If the learning experience that goes along with the growth of awareness does not speak to these logical connections, then the agent will not have enough information to form a unique posterior credence. This is not a feature of my proposal for handling expert testimony however, it is a feature of the problem of awareness growth.

2.6 DISAGREEING EXPERTS

We now return to the problem of Expert Disagreement. As we saw in the discussion of Unconditional Deference, the belief revision schema approach has not yet said anything about how to deal with the general *n*-expert case.

Let me start by distinguishing between two kinds of expert disagreement. The first is an easy case in which the expert reports, while different, are compatible. Consider one expert who says that a probability is above 0.5, while another says that it is below 0.6. There is a set of probability functions that is compatible with both: those that say the relevant probability is in the interval $(0.5, 0.6)$.

On the other hand, if the first expert said it was *below* 0.5, and the second that it was *above* 0.6, then there would be no probability functions compatible with both reports.

My approach has a natural way of dealing with "easy" cases. Recall that an expert report delivers a constraint, ϕ_Q, or equivalently determines a set of belief states that is compatible with it. It is natural to think that multiple expert reports jointly determine an input: the conjunction of their individual constraints. The agent can defer to this joint input, taking it as a perturbation of their belief state.

Following the recipe above, we can now look for propagation procedures. But we have a choice at this stage. Our first option is to stick with a precise probability model, in which the agent must have a single probability for the proposition, lying within the range $(0.5, 0.6)$.

The second option is to allow for imprecise probabilities, and have the agent adopt the range $(0.5, 0.6)$ as their belief state.

If we stick with a precise model, we will need some principle for choosing from the allowed range. There are many possibilities. Perhaps if the agent's prior is within the interval, then they can retain it. For the moment, however, let's suppose that all options are permissible. The agent then picks one and uses an appropriate propagation procedure to bring their credences to coherence. If the content of the report is a familiar proposition, then the agent proceeds as described in section 2.4. If it is an unfamiliar proposition, they proceed as described in section 2.5.

If we go with an imprecise model, then we are in the same situation discussed briefly at the end of section 2.4, where I considered imprecise reports. There is room for disagreement here about how to conduct the propagation step. One natural approach is to generate an imprecise posterior state by performing the propagation using every permissible function in the input. Suppose that we are in an Unconditional Deference case, and the experts have reported on the probability of rain. The perturbation sets your imprecise probability for rain to $[0.2, 0.4]$. You now propagate this through your belief state by Jeffrey conditioning. But instead of doing it once, you do it for every value in the interval $[0.2, 0.4]$. In this way you generate an imprecise posterior credence.

If the expert reports are not compatible, expert deference cannot help us without further supplementation. In the next chapter, I discuss Opinion Pooling, which is a more general attempt to solve the problem of disagreement. If we endorse some form of pooling, then the result of a pooling procedure might itself be a good candidate for expert deference.

2.6.1 *Conclusion*

My investigation of Bayesian approaches to expert testimony has led me to develop a new model of expert deference. My proposal was constructed to deal with the problems identified for supra-Bayesianism and expert deference as a constraint on priors. The problems with supra-Bayesianism were: Cognitive Burden, Relevance of Priors, Sensitivity to Testimony, and Awareness. The orthodox model of deference did better on the first three, but introduced two additional problems: Arbitrariness, and Expert Disagreement.

To review: the problems for supra-Bayesianism were associated with the propositions representing the experts' reports, and the content of those reports. I therefore removed the expert reports from the algebra entirely, instead of representing them as propositions they are now externally given constraints on the agent's posterior beliefs. The Bayesian updating procedure has been replaced with expert deference (now realised as the imposition of this external constraint) and a belief revision schema, in which Bayesian conditioning is one element. This reduces the cognitive burden on agents, as they are not required to have a myriad of prior beliefs. It does not depend on the agent's uninformed priors for the propositions in the expert domain, as those play no role (and aren't required to exist). In deferring, it is properly sensitive to the content of the expert's testimony.

I depart from strict Bayesian orthodoxy by considering a wider range of updating rules. This allows me to avoid arbitrarily treating only expert reports of unconditional probability as worthy of deference. My proposal will, therefore, inherit concerns about, for example, Jeffrey conditioning, but I do not think that it introduces additional concerns. As I have indicated for the case of Jeffrey conditioning, I think that those concerns can be dissolved with proper analysis, and so I am content with it.

Two problems remain: Awareness and Expert Disagreement.

I have shown how my proposal fits naturally with one proposal for rational awareness growth, Bradley's "rigid extension". Cases involving awareness growth are difficult, often leading to very poorly constrained imprecise credences. But this is not specific to my proposal (about expert testimony), it is a feature of awareness growth as a phenomenon. Indeed deference provides a simple way of setting attitudes on the new algebra that would otherwise be difficult to justify.

What of disagreement? Expert deference as a constraint on priors lacked the resources to deal with disagreement because it is too blunt a tool. It has no way of handling multiple instances of expert testimony. My proposal offers a slight improvement: when the testimonies are compatible, it offers a natural way of deferring to the common ground between the disagreeing experts. But in more severe cases of disagreement it doesn't offer us much.

I take this chapter to demonstrate that Bayesian methods can be of use in studying expert testimony for realistic agents. What is required is a willingness to move beyond Bayesian orthodoxy and adapt Bayesian methods to reflect the limitations of the agents being studied. However, the problem of most interest to me, expert disagreement, is the one where least progress has been made. In the next chapter, I therefore turn to other formal models of disagreement.

2.7 ADDENDUM: BELIEF REVISION BY DIVERGENCE MINIMISATION

In section 2.4, I noted that there is no general answer to the question of how to fix a unique posterior given just a perturbation. The kinds of expert deference (unconditional, conditional, experiential) that I reviewed in section 2.4 as instances of my belief revision schema are particular cases with very strong conservation conditions which allow us to identify a unique kinematic update rule which generates a unique posterior. In this addendum, I consider one approach to answering the more general problem.

To start, let's consider a case involving an unconditional probability. Suppose the agent hears expert i report their credence for X. Imagine the processes described taking place in a sequence, and say that the agent has just adopted the reported credence as their own, but that they have done nothing else so far. We can describe their partial belief state with:

$$P^*(Y) = \begin{cases} P(Y), & Y \neq X \\ P^i(Y), & Y = X \end{cases}$$

This is not a probability function: it is not additive for disjunctions involving X. What is needed is a method for bringing it to coherence; or, put another way, for using P^* to find a coherent probability function. How might we go about finding the right probability function, without using Rigidity—or, if you like, if we're in a situation where Rigidity doesn't hold?

In the absence of explicit conditions on which parts of the prior function should be preserved, one way that philosophers have attempted to model this is by looking for a probability function that is closest to P^* while being responsive to the experience, under some definition of "closeness." I will now consider the prospect of using such a "distance-based" approach here.[20]

The tools that mathematicians have developed for measuring the difference between two probability functions are called *divergences*. Suppose we have some algebra, and we form the set of all probability functions defined on it, Π. A divergence is a function $\mathcal{D} : \Pi \times \Pi \to [0, \infty]$, such that (i) $\mathcal{D}(P, P) = 0$ for all $P \in \Pi$, and (ii) $\mathcal{D}(P, Q) > 0$ for all $P \neq Q$.

[20] There is a large literature on using "distances" between credences for epistemological or decision theoretic purposes: such methods are present in foundational work by de Finetti and Savage; two sources I have already made significant use of—van Fraassen (1981) and Diaconis and Zabell (1982)—discuss and utilise such methods; and they are present throughout the "accuracy" programme in epistemology including Joyce (1998), Leitgeb and Pettigrew (2010b), and Pettigrew (2016).

Divergences are used to quantify differences but they aren't distances, or *metrics*, as mathematicians term the functions that measure distance. A metric is symmetric ($d(P, Q) = d(Q, P)$) and obeys the triangle inequality ($d(P, R) \leq d(P, Q) + d(Q, R)$). Divergences do not need to obey these properties. So, while I will use adjectives like "close" in my informal discussion, it is important to remember that divergences *are not distances* as we know them.

So what are they? To ground the discussion, I will start with two common examples of divergences. Consider first the Kullback-Leibler divergence, popular in information theory.

Definition. (Kullback-Leibler divergence)

$$KL(P, Q) = \sum_{X \in \Omega} P(X) \log \left(\frac{P(X)}{Q(X)} \right)$$

As you can see, it takes the form of a weighted average. For each proposition, we form the ratio of the probabilities assigned to that proposition, then take the log of that ratio and weight it with the probability that the *first* probability function assigns to the proposition. I won't discuss why this is a useful measure of difference, but will point out that it is obviously not symmetric: the order of P and Q in the arguments matters.

A second example is more familiar: the Squared Euclidean divergence.

Definition. (Squared Euclidean divergence)

$$S(P, Q) = \sum_{X \in \Omega} \Big[P(X) - Q(X) \Big]^2$$

As the name implies, this divergence is closely related to our notion of distance in Euclidean space. It is symmetric: the fact that the difference is squared means that the order of P and Q doesn't matter.

Proposal: The proposal that I want to consider for using these functions in a belief revision rule is as follows: we find the probability function Q, such that the divergence between Q and P^* is minimal.

But, as we have noticed, divergences are not generally symmetric. There are therefore two potential candidates for the revised function, resulting from minimising the left divergence and right divergence from P^* respectively.

$$Q_{L\mathcal{D}}(P^*) = \operatorname*{argmin}_{Q' \in \Pi} \mathcal{D}(Q', P^*) \tag{4}$$

$$Q_{R\mathcal{D}}(P^*) = \operatorname*{argmin}_{Q' \in \Pi} \mathcal{D}(P^*, Q') \tag{5}$$

In addition, there are a great many divergences. I have shown two examples, but the class of divergences is infinite and even more restricted classes are very large. In general, the divergence-minimisers for two different divergences will be different.

Indeed, we should not expect any of these to agree: the left- and right-divergence minimisers can be different, and different divergences will generate different (left- and right-)divergence-minimising functions.

What is needed to fill out this approach is an analysis that selects a class of divergences that are suited to our task. Ideally, we will be able to find a unique divergence for belief revision, which will return the accepted answers to familiar cases (certainly Bayesian conditioning and, many would hope, Jeffrey/Adams conditioning for the appropriate experiences).

How might we undertake such an analysis? What is required is a philosophical motivation for various properties that restrict the class of permissible divergences. For example, the fact that eqs 4 and 5 do not agree in general would be resolved if we considered only symmetric divergences, for which $Q_{L\mathcal{D}} = Q_{R\mathcal{D}}$. But how can we justify such a restriction? Richard Pettigrew provides such an analysis in his discussion of why we should use a symmetric divergence in our definition of an accuracy measure for credences. Here is the relevant passage:

> *We have a strong intuition that the inaccuracy of an agent's credence function at a world is the distance between that credence function and the ideal credence function at that world. But we have no strong intuition that this distance must be the distance from the ideal credence function to the agent's credence function rather than the distance to the ideal credence function from the agent's credence function; nor have we a strong intuition that it is the latter rather than the former. But if there were non-symmetric divergences that gave rise to measures of inaccuracy, we would expect that we would have intuitions about this latter question, since, for at least some accounts of the ideal credence function at a world and for some agents, this would make a difference to the inaccuracies to which such a divergence gives rise. Thus, there cannot be such divergences. Symmetry follows.* (Pettigrew, 2016, p. 67)

The strategy here is to take intuitions about good partial beliefs, and to use them to determine the right kind of mathematical representation, in this case for the relevant notion of "goodness" itself. The basic intuition being appealed to is that we want to be "close" to the truth (the ideal credence function assigns 1 to true propositions and 0 to false propositions). This has some appeal: our beliefs can be wrong in different ways, and we commonly talk about some being worse than others despite all being false. Maybe I believe it is sunny and warm outside, while you believe it is sunny and cold. In fact, it is overcast and cold. There's a sense in which your belief is better than mine though we were both wrong.

But then Pettigrew invokes the *lack* of an intuition as a reason to restrict the class of permissible divergences. It is because we have no intuition about the direction of the "distance" that we demand that the divergence in question be symmetric. But here I worry that the use of the word "distance" and our intuitive familiarity with distances is doing too much work.

Consider the intuition of *conservatism* of belief revision: we want to remain close to P (the initial credence function) when we identify Q (the final credence function). P is a fixed point, while Q is a variable—so it makes little sense to speak of remaining close to Q. This establishes some difference between them, some priority for P.[21] Now as soon as we think about the "distance between" the two functions, it

[21] Similarly in the accuracy case, there is a difference of role between our credences and the truth, arising from direction of fit. My credences must be close to the truth, rather than the truth needing to be close to my credences. Once we speak only of "distance" this intuition is lost.

is hard to sustain that priority, but this *is* because our intuitive notion of distance is symmetric.

If we start from the position that divergences are the right tools to measure the difference between probability measures, however, then we already accept that we *must not* think in terms of distance. (I'm not trying to exclude symmetry from the get-go, but to rely on intuitions about symmetric distance is to beg the question.) Indeed, the lack of symmetry can be a way for us to reflect the different statuses of P and Q, one fixed and the other variable. I doubt that we can find a philosophical argument for which of the two arguments of $\mathcal{D}(\cdot|\cdot)$ ought to be used for the fixed point, and which for the variable. Instead we face a conventional choice between two formalisms: a left-variable formalism in which $Q_{L\mathcal{D}}$ is the correct definition, and a right-variable formalism in which it is $Q_{R\mathcal{D}}$. For example, Eva, Hartmann, and Rad (2019) do not discuss the lack of symmetry problem, but they appear to adopt the convention that one always places the posterior in the first position in the divergence.

Let us return to equations eqs 4 and 5. My presentation above contained a misleading simplification. As written, eqs 4, 5 will not produce the right revised credence functions, for we have no guarantee that the closest function to P^* won't be P—i.e., that our divergence-minimising procedure will not *undo* the update due to deference. So we need to restrict the domain for argmin to those probability functions which preserve that update: $\Pi^* = \{Q \in \Pi : Q(X) = x^i\}$. The revision procedure is a constrained optimisation problem, where one constraint is supplied by the expert testimony, as discussed above.

The problem we face is that these constraints simply tell us how to carry out the minimisation, *whichever* divergence we choose. They give us no grip on the problem of divergence choice.

There seem to be two options available to us. We can either motivate for a divergence by some other means, and then apply that favoured divergence to the problem of belief revision. Or we can attempt to apply the results of belief revision to narrowing down the set of divergences.

Here is an example of the first approach. Leitgeb and Pettigrew (2010a) provide an accuracy-based proof for Probabilism and updating by Bayesian conditioning for Bayes inputs. They do this by arguing for a particular inaccuracy measure—a generalisation of the Squared Euclidean divergence—and a particular norm for minimising inaccuracy. They then go on to show (in Leitgeb and Pettigrew, 2010b) that Jeffrey conditioning does not agree with their procedure in general: i.e., that minimising inaccuracy after a Jeffrey-type input using the SE divergence does not agree with the result of Jeffrey conditioning. Rather, following their recipe is equivalent to a revision procedure in which one adds a constant to the probability of all worlds in which X is true—a procedure dubbed LP-updating by Levinstein (2012). For my purposes, the important thing here is the method of argument: Leitgeb and Pettigrew had already identified the "right divergence" in their derivation of the core Bayesian norms, and they were then applying it to the particular case of updating on Jeffrey inputs—with no other constraints.

The problem is that their updating rule has some highly counterintuitive consequences. Ben Levinstein (2012) demonstrated that LP-updating alters the evidential relationships encoded in the prior probability function; it fails to preserve the ratios of likelihoods between propositions that existed in P. If an agent considers $\mathbb{X}$ and $\mathbb{Y}$ to be independent, and learns new probabilities for $\mathbb{X}$ (through evidence that doesn't bear on this dependency), an LP-update can nevertheless radically alter the probabilities of $\mathbb{Y}$. Levinstein gives an example in which an agent becomes more certain that a car is red, and this leads them to radically increase their credence in the existence of ghosts, despite holding these propositions to be independent (Levinstein, 2012, p. 420).

This is, in a sense, a direct consequence of LP-updating not obeying Rigidity, the condition that Jeffrey proposed as the right way for belief revision to be Conservative in the face of Jeffrey inputs. If Rigidity is imposed as a constraint on which posteriors are permissible, then minimising inaccuracy using the SE divergence does yield Jeffrey conditioning. Now, Leitgeb and Pettigrew (2010b) were well aware that their rule did not obey that Rigidity condition. But they were engaged in an axiomatic programme, aimed at deriving belief revision norms from more basic considerations. Given their project, it would not be legitimate to impose Rigidity as an additional constraint.[22] As Levinstein puts it: "such a move looks like an ad hoc fix unless more motivation can be provided…Structural requirements

[22] The Rigidity condition for Bayesian conditioning is a consequence, in their system, rather than a constraint (Leitgeb and Pettigrew, 2010b, p. 262).

on a credence function should emerge from evidential and alethic requirements" (2012, p. 53).

We need not adopt these goals, however. My purpose is to find a reasonable updating procedure for an agent who wishes to defer to expert reports in a wide range of cases. Those cases include situations in which the various Rigidity conditions discussed above do hold. Could we not, therefore, reverse the order of justification? We have cases where there is a widely accepted right answer for the belief revision. For Bayes-type inputs, the revision should match the result of updating by Bayesian conditioning. For Jeffrey-type inputs, it should match the results of Jeffrey conditioning. And for those who accept Bradley's Adams conditioning, the same thought applies. Can we select a divergence by insisting that it obey known Conservation conditions for familiar input types? The answer is yes, but this robs divergence-based reasoning of any real interest.

Suppose that we are considering our unconditional deference case, and we're looking for an appropriate divergence. We note that the posterior must have $Q(X) = x^i$, where x^i is the expert's reported credence, and that we antecedently take the right answer to be given by Jeffrey conditioning. We search the space of divergences, and find that there *are* divergences which reproduce Jeffrey conditioning. One class of them is the class of f-divergences (Eva, Hartmann, and Rad, 2019, Proposition 3).

Definition. (f-Divergence) Let f be a convex function such that $f(1) = 0$. Then for any such f, the f-divergence between Q and P is:

$$\mathcal{D}_f(Q, P) = \sum_{X \in \Omega} P(X) f\left(\frac{Q(X)}{P(X)}\right)$$

We have already seen an example of an f-divergence: the Kullback-Leibler divergence, for which $f = x \log x$. This gives some succour to the divergence minimiser, but we're left with a very large class of possible divergences. We could perhaps refine it further. If a motivation for symmetry can be found, there are symmetric f-divergences, such as Symmetrised KL: $KL(P, Q) + KL(Q, P)$. Or we could look for divergences which also reproduce the results of Adams conditioning, or Odds kinematics. But this puts the divergence-minimiser in an invidious position. As Jeffrey put it in 1983: "what we thereby discover is that…we have adequate concepts of closeness" for we already knew that Bayesian and Jeffrey conditioning were the rules appropriate to their learning experiences (Jeffrey, 1992, p. 81). It is unclear what value is added to our project by reframing familiar conditions in terms of divergences.

There isn't much room for hope in shifting attention to the Conservatism conditions. One might think that, rather than straightforwardly insisting that we recover Jeffrey conditioning, we should instead insist that the chosen divergence respects the relevant Conservatism condition. But in the cases discussed above, the conjunction of Responsiveness and Conservatism uniquely selects the appropriate update rule: Bayesian, Jeffrey, or Adams conditioning.

To my mind, the conclusion is this: how to revise belief depends on the kind of evidence one is revising in light of. Learning experiences change some parts of one's initial beliefs and leave others untouched. Paying attention to these gives us traction in dealing with kinds of learning that cover wide ranges of potential learning experiences, including many cases of expert testimony. While there is great theoretical interest in pursuing a foundational project in epistemology that subsumes belief revision under a broader—perhaps accuracy-based—approach, I do not see good immediate prospects for adopting that approach for my purposes here.

3

OPINION POOLING

3.1 INTRODUCTION

A very common answer to the problem of expert disagreement is to aggregate, or pool, the views of the panel. Recall Case 1:

Case 1. *Ade is a policymaker, trying to decide how to enhance Thames flood defences for the next fifty years. He wishes to use the best scientific advice available to determine the likelihood that the Thames will rise more than 50cm—which would require new barriers. He convenes a panel of experts. The 10 experts disagree, offering a wide range of answers, from unlikely to very likely.*

Given Ade's epistemic position, he cannot choose one expert to follow nor adjudicate the debate. But there is some evidence that Ade can respond to: the *distribution* of expert answers. The aggregation approach assumes that some facts about this distribution can guide Ade toward a rational resolution of the disagreement. In most cases, "aggregation" means using some central tendency of the distribution of expert reports as the single opinion for decision-making or belief formation; for example, the simple linear average of the expert reports. This is a popular approach to the problem of expert disagreement, and forms the basis of many expert elicitation procedures (e.g., Cooke, 1999, 2018 and Aspinall, 2010). Averaging also shows up in the philosophical literature on disagreement between epistemic peers, where defenders of the "equal-weight" view argue that it is the best response to such disagreements (e.g., Elga, 2007 and Christensen, 2007).

Supposing that we wish to average opinions, we must now acknowledge that there are different ways to average. To name the most popular in the philosophical and statistical literature: linear, geometric, and multiplicative. In much of the literature on opinion pooling the discussion of these methods takes place in the abstract (i.e., without reference to a particular problem) because pooling could be used for a number of different purposes. One might be interested in the construction of a group agent, or the design of a machine learning algorithm, or the resolution of a peer disagreement problem. This literature therefore offers characterisations of different pooling procedures, with the implication being that the context of use would determine which characteristics are desirable, and therefore which pooling procedure to select. This chapter offers such a contextual analysis for expert disagreement, evaluating linear and geometric pooling as options for solving the problem of expert disagreement as it appears in cases like Ade's. (Multiplicative pooling will prove to be covered by my discussion of geometric pooling.)

This chapter's discussion of aggregation procedures will make use of the language and formalism of probabilistic opinion pooling (as constructed by Dietrich and List, 2017). In this literature, opinions are represented by probability functions, and so the most natural application is to cases in which the expert reports are themselves the probabilities of events. However it can be naturally extended to reports that are about the values of variables by construing those reports as offering means (if they are point valued) or confidence intervals (if they are range valued) of the distribution of a random variable. While the formal framework is very similar to that of chapter 2, we should not assume that we are in a strict Bayesian setting. We are working with probabilities, and will look for links to the Bayesian theory of rational degrees of belief, but that theory is no longer assumed.

Below, reported opinions are denoted P^i where i ranges up to n the number of experts. Opinions are about propositions, $X, Y, \ldots$ collected in an agenda $\mathcal{X}$. Opinions are aggregated by a *pooling function*, denoted F, which takes the n opinions as inputs and produces a single opinion function P. The three kinds of averaging mentioned above are here conceived of as different forms for the pooling function:[1]

- **Linear pooling**: $F(P^1, \ldots, P^n) = P = \sum_i w_i P^i$, where $\sum_i w_i = 1, w_i \geq 0 \; \forall i$.
- **Geometric pooling**: $F(P^1, \ldots, P^n) = P = c \prod_i {P^i}^{w_i}$, where $\sum_i w_i = 1, w_i \geq 0 \; \forall i$, and c is a normalisation factor.
- **Multiplicative pooling**: $F(P^1, \ldots, P^n) = P = c \prod_{i=0}^{n} P^i$, where P^0 is a calibrating function, and c is again a normalisation factor.

Geometric and multiplicative pooling are defined by pooling the probabilities of "worlds" rather than propositions. These can be thought of either as maximally specific propositions (roughly, non-contradictory

[1] See Dietrich and List (2016a) for a review; Dietrich (2010a) is the source of multiplicative pooling in the philosophical literature.

conjunctions involving as many propositions as possible) or as elements of an underlying sample space on which the propositions are defined as subsets. Geometric and multiplicative pooling is easily extended to propositions by forming sets of the relevant worlds. Dietrich and List (2016a) provide details. As the notation indicates, the calibrating function in multiplicative pooling is a probability function. It is related to the individuals' priors, in a manner outlined in Dietrich and List (2016a, S.9).

Each explains how to combine individual opinions. Linear pooling says that for any issue X, the aggregate opinion is formed by multiplying each individual i's opinion on X by a weight w_i, and then adding these weighted opinions together. Geometric pooling advises us to instead raise each opinion to the power of a suitable weight w_i, and then to multiply them together. Multiplicative pooling involves multiplying the opinions together, and then multiplying by a calibrating function P^0 that is a weighted average of the experts' priors. Note that in each case the pooling function requires *weights*, which represent a judgement of how much say each individual should have in the aggregate opinion. The choice of pooling function does not tell us how to set these weights; we will turn to that question in section 3.4.

3.2 LINEAR POOLING

I'll start the discussion with linear pooling then move on to geometric and multiplicative. I begin by examining whether there is direct motivation for using the linear average of expert reports. I will then turn to one of the major ways this question has been addressed in the statistics and philosophy literature: axiomatic characterisation of pooling functions. Here I will examine how an agent in a realistic case where pooling is to be applied would select a function using axiomatic considerations.

3.2.1 *General motivation*

For all that linear averaging is a popular approach to expert disagreement,[2] it is rare to find much in the way of direct philosophical motivation for it. I will therefore present various arguments for averaging that I have encountered, accompanied by reflections on what they establish.

3.2.1.1 *Peer disagreement*

One thought that might come naturally to philosophers is that the literature on *peer* disagreement should be important to our discussion of expert disagreement. The experts, we might suppose, are peers to one another; they disagree; and so they are bound by whatever norms govern peer disagreement. If they fail to enact the required changes on their own we might consider how to enact them ourselves.

One of the popular views in this literature is called the "equal weight view" which, as the name indicates, argues that we should

[2] It *is* popular: Clemen (1989) provides an annotated bibliography of 200 studies just on the aggregation of forecasts, and notes that there is a large literature on aggregating other kinds of models in economics.

give equal weight to our own and our peers' views (e.g., Elga, 2007). In the credal case, this appears to amount to adopting a simple average of all the reported credences (including one's own reported prior) (Jehle and Fitelson, 2009; Kelly, 2010).

While this is an appealing idea, there are a number of problems with it. First, the peer disagreement debate uses a highly specific and idealised notion of peerhood—or rather, several competing such notions. Peers have access to a common body of relevant evidence, and no other relevant evidence; they are equally good at evaluating the claim at hand (Elga, 2007, p. 484). Peers have had long discussions in which they share everything they can think of that is relevant to their disagreement; they are equally intelligent and rational and it is known that neither is in a state that would undermine their ability to evaluate the claim at hand (Christensen, 2007, p. 188). Peers have the same evidence, and indeed have made the same observations. To the extent that track records are available, they indicate equal reliability for peers (Kelly, 2010). Among these three examples there are common strands (common evidence, equal skill) but there are also subtle differences.

Now how do these apply to the expert case? In our case, we might not want to assume equal evidence. Though we can ask experts to share data, this will not accord with the high standards of common evidence found in the disagreement literature. Experts will have had many experiences over the course of their lives the contents of which they cannot share in any simple manner. They may not recall everything that they're using to make their judgement. And, more prosaically many expert elicitation procedures operate under time pressures that block investigation, sharing and processing of private evidence. Similarly, we may be able to bring about some deliberation but may also wish for a solution that works in the absence of sufficient deliberation as defined by epistemologists. Track records will feature heavily in section 3.4, and so I don't wish to build equality into the definition of expert disagreement here.

From the formal epistemologist's perspective, a final issue is how we *identify* peers given a specification of a group of agents' credence functions. While this issue is somewhat more of an ideal theory concern than my usual fare here, I will briefly sketch the worry. To start, there is little by way of formal definition of peerhood in the peer disagreement literature. Where such definitions have been offered they vary even more widely than the non-formal definitions. Jeffrey (1992): equally good testimony should be defined in terms of reliability, here thought of as the probability of true and false positive reports. Elga (2007, p. 487): "conditional on a disagreement arising, the two of you are equally likely to be mistaken." Joyce (2007): “epistemic comrades” will agree on certainties, agree in expectation, and will expect no disagreement amongst two other peers. Easwaran et al. (2016): an agent thinks that their peer is more likely to have higher credences in X when X is true; perhaps by having likelihoods for their peer's credences which are linear in those credences.

Finally, there is not much focus in the peer disagreement literature on the question of the precise mathematical form that averaging should take; it is largely assumed that the averaging should be linear.

While there is some nascent discussion of other forms of averaging emerging in that literature it seems to me that, given the differences between the expert and peer case, those of us interested in expert disagreement may as well perform the analysis anew ourselves. In any case, for the purposes of this section, it does not seem that we can rely on the peer disagreement literature as a *motivation for* linear pooling in the expert disagreement case.

3.2.1.2 *(Approximate) Bayesianism*

The orthodox Bayesian response to expert disagreement was discussed in chapter 2 under the name supra-Bayesianism. But, as many have noted, this is an extremely difficult procedure to put into effect in actual cases of expert disagreement (Genest and Zidek, 1986). A natural question to ask is whether there is a (simpler) procedure we can implement that will "match" the results of the supra-Bayesian procedure—a condition that is known as Bayes-compatibility. I will not try to address this question in general, but will consider whether some form of averaging could be such a procedure.

It is easy to see that averaging cannot *precisely* match the behaviour of Bayesian conditioning. For one thing, Bayesian conditioning is commutative and associative; while averaging is not. More generally, conditioning and averaging behave quite differently with respect to the underlying information encoded in the reported probabilities. (Averaging, after all, is a procedure that can be carried out with any numbers while conditioning is a peculiarly probabilistic operation.) Bradley (2007) presents a simple case to show how this can lead averaging to the wrong answer. First, note that in the standard presentation of opinion pooling the weights depend on the expert but *not* on the proposition (see Dietrich and List, 2016a; Genest and Zidek, 1986). (This is because we are aggregating probability functions on a Boolean algebra, and a requirement of the procedure is that the aggregate opinion is also a probability on this algebra. If each expert report received a different weight for each proposition, the aggregate function could not be both additive and normalised.) However, if this is the case, any averaging procedure will be in tension with the common-sense demands of rationality.

Bradley (2007) shows this: suppose that person 1 observes A and person 2 observes B. Before consulting, they each had the joint distributions shown in Table 3. They then learn of each other's observations. Trusting one another completely, it is clear that they should now agree on the distributions in Table 4.

Table 3: Prior beliefs

	AB	$A\neg B$	$\neg AB$	$\neg A\neg B$
P^1	0.3	0.7	0	0
P^2	0.3	0	0.7	0

Table 4: Posterior beliefs

	AB	$A\neg B$	$\neg AB$	$\neg A\neg B$
Q^1	1	0	0	0
Q^2	1	0	0	0

But this cannot be the outcome of any averaging procedure that is independent of the proposition. Only by adopting two sets of weights, one which allocates full confidence on the question of A to person 1, and another which allocates full confidence in B to person 2, could Table 4 be achieved via averaging.

We should note here that supra-Bayesianism will only recover this result under conditions of full trust, which is to say, deference. Bradley therefore considers whether linear pooling can be Bayes-compatible in expert deference cases. Taking into account that people are experts in particular domains only, we now consider proposition-dependent weighting and deference in their topic area only. We ask whether this procedure can be matched by pooling the expert reports within their areas (rather than pooling their full credence functions).

Bradley (2018, p. 7) argues that “Linear averaging, even when applied to a single proposition, is not (non-trivially) compatible with Bayesian conditionalisation in situations involving more than one source of expert testimony to which deference is mandated.” If averaging is to be compatible with deference, and the weights are to be independent of the reports made, the experts must always make the *same report*—in which case, the case for consulting more than one expert disappears (Bradley, 2018, pp. 16–17).

Expert deference, however, is a special case. Can we say anything about Bayes-compatibility in general? Genest and Schervish (1985) establish that we can. If we are free to choose the agent’s likelihoods, then it is always possible to ensure that the supra-Bayesian result matches that of linear pooling for some choice of likelihood and expert weights. Jan-Willem Romeijn (manuscript) has put these into an especially easy to digest form as follows. Suppose once again that our agent has probabilities P, and expert P^1, and that the proposition at issue is X. Let the agent's prior be $P(X) = x$, and the expert's $P^1(X) = x^1$, and let $\ulcorner x^1 \urcorner$ be the proposition that the expert reports this probability. The supra-Bayesian updates to $P(X|\ulcorner x^1 \urcorner)$, while the result of linear pooling is $Q(X) = wx^1 + (1-w)x$. Romeijn (manuscript, p. 4) reports a corollary of Genest and Schervish's result: $Q(X) = P(X|\ulcorner x^1 \urcorner)$ if we choose the likelihoods

$$\begin{aligned} P(\ulcorner x^1 \urcorner|X) &= g(x^1, x)\left(1 - w + w\frac{x^1}{x}\right) \\ P(\ulcorner x^1 \urcorner|\neg X) &= g(x^1, x)\left(1 + w\frac{x}{1-x} - w\frac{x^1}{1-x}\right) \end{aligned}$$

where $g(x^1, x)$ is the agent's unconditional prior for the expert making this report, denoted in this way to highlight the importance of the agent's prior in X, i.e., x, and which must be such that

$$\int_0^1 g(x^1, x)dx_1 = 1; \qquad \int_0^1 x_1 g(x^1, x)dx_1 = x.$$

Note that these integrals are with respect to x^1, the expert's report. It is natural to have this constraint, which ensures that the agent's prior for the expert's report centres on the agent's own prior x.

The problem is that this is not of much use to either the Bayesian epistemologists or the policymaker facing expert disagreement. It puts the cart before the horse, telling us how to ensure that supra-Bayesianism matches a particular linear pooling result. But to a Bayesian epistemologist it is supra-Bayesianism that is the normative ideal, and the likelihoods *aren't* free parameters, they are elements of the agent's attitude of partial belief. For a given prior partial belief state, and its associated particular supra-Bayesian posterior upon hearing a report, there is *no reason* to expect agreement with linear pooling. But perhaps there is some comfort offered to the policymaker using linear pooling? After all *some kind* of Bayesian could have gotten this result. No: *some* Bayesian could achieve any result if enough of their prior is taken as a free parameter. This result does not guarantee that the *policymaker's* particular action is rational, given *their priors*. So, once again, we find no support for linear pooling in the expert disagreement case.

3.2.1.3 *Sampling*

Recall some basic statistics. There is a large set of objects of interest to us, called a population. We would like to learn some properties of this population, but as it has so many members it is impractical to take a census in order to completely determine these properties. A sample is a set of data collected from a population, on the basis of which we will make inferences about the population. Samples are typically selected according to some properties of the members of the population, and a representative sample is one chosen using a selection process that does not depend on other properties of the population. For example, a representative sample of English voters in the 2017 election might consist of a randomly sampled set of 10,000 of the English people who voted in the 2017 election. On the other hand, a sample chosen from English Twitter users who voted in the 2017 election may not be unbiased, since many English voters are not on Twitter.

An *estimator* is a statistic (a function of the sample data) that produces an estimate of a desired quantity. A random sample is one in which every element of the population has a non-zero probability of being selected as a member of the sample, according to a probability measure on the population that is either known or can be determined. Such samples make it possible to produce unbiased estimates of population properties, as we can weigh the properties of elements of the sample according to their probability of selection to be in the sample. The sample mean of a random sample is an estimator of the mean of the population. It is an unbiased estimator, as the expected value of the sample mean is the population mean.

Here is one motivation for averaging expert opinions, that is often gestured towards or mentioned in conversation, but which I have not found defended explicitly in print. Expert elicitation is analogous to measurement (itself a form of sampling), in the following way. There is some phenomenon of interest, which motivates the formation of the expert panel. It determines some variable or variables of interest. Expert reports are like measurements of these variables, generating members of a sample. (Perhaps because the experts are scientists who do take such measurements and whose reports will be based on these, amongst other things.) Weights given to expert reports in the pooling procedure are like the weights of the probability distribution on the population determining the sample. Thus the average of the reports is a sample mean: an unbiased estimator of the population mean. The population mean, in the analogy, takes the role of the true value for the variable in the population.

The analogy does not secure the use of the mean report however. We are not in fact measuring some phenomenon; we are eliciting judgements from experts in a manner that does not conform to the requirements of random sampling. The weights used in pooling typically reflect the presumed credibility or skill of the experts, not their probability of selection in a sampling process. So there is no reason to believe the average of the reports is an unbiased estimator of some population mean, nor is there a clear sense of what that population mean would be or why we want to estimate it.

Here's a second motivation: expert panels *are* samples. We selected experts to be on our panel from a population of experts. So we can think of our set of expert reports as a sample taken from the population of reports that the population of experts would have produced. We wish to determine the mean report in the population of expert reports. We use the sample mean as an estimator.

The first problem here has already been mentioned: the weights assigned to experts are credibility weights rather than sampling probability weights. Secondly, there are worries about whether we should want to find the average expert report. Goldman (2001) argues that the agreement of other experts with one is not reliable evidence that the one is correct. Goldman's arguments concern the independence of the additional experts, in a statistical sense. Suppose Dayo is supported by experts Ige, Ojo, and Ale. If Ige is a "slavish follower" of Dayo, and merely agrees with her report *because* Dayo says so, then we should not allocate any weight to their position. (In Bayesian language: if we have already conditionalised on Dayo's testimony and so formed the credence $P(H|D)$, then if the additional expert reports D^i's are entirely dependent on D, $P(H|D \cap_i D^i) = P(H|D)$.) In order to provide additional support, the additional experts' opinions must be at least partly conditionally independent of the existing experts (Jeffrey, 1992, pp. 108–10).

Goldman acknowledges that, in the case of scientists, we will typically have reason to expect them to be critical of one another's positions. But to the extent that there is dependence, the average will not track the epistemically relevant quantity. So, without further qualification, we have no reason to use the linear average.

3.2.1.4 *(Other) convergence arguments*

More generally, one might appeal to to mathematically similar convergence results such as Condorcet's Jury Theorem.

The theorem works like so: suppose we are deciding on the truth of a proposition, and we have a set of voters, each of whom is independently >50% likely to get the truth-value correct. We count votes and determine the majority position. The Theorem says that the larger the number of votes, the higher the probability that the majority is correct, and as the number of voters increases that probability tends to 1. (I note at the outset that this is a result about the truth of a proposition, rather than estimates of its probability, and so it cannot apply *directly* to our case. However some authors have found connections between pooling and Condorcet-type situations, e.g. Romeijn (manuscript).)

The Jury Theorem has some important limitations: real voters are rarely independent, and in cases where their competence is low (close to 0.5) and correlations between them are high, their collective competence can fall below that of a single juror (Kaniovski and Zaigraev, 2011). In the propositional setting the competence assumption is generally glossed as a requirement that the experts be better than chance. When the propositions are of the form $\ulcorner \P(X) = x \urcorner$ then the "doing better than chance" gloss will not do: it would mean only having nonzero competence, as the set of such propositions that chance draws from is infinite even for fixed X.

Independence is a difficult issue for expert panels. Experts are surely not going to be unconditionally probabilistically independent—their expertise in a common domain means they will share a set of tools, theoretical assumptions and biases. They will be familiar with a set of canonical problems, were trained on the same historical data, and so forth (Dietrich and Spiekermann, 2013, p. 10).

To deal with this, Dietrich and Spiekermann (2013) develop a theorem that depends on a new form of conditional independence, where what is conditioned on is a "problem"—which captures facts about the state of the world, as well as common causes between voters. Experts will satisfy this new independence criterion. But with it comes a conditional competence assumption: experts must have problem-specific competence (>0.5 probability of a correct judgement). This is much harder for us to assume in general, as laypeople. We know that when experts face difficult problems, such as economists predicting the 2008 financial crisis, many of them can fail. The new theorem secures independence, but at the cost of making us significantly less certain of expert competence. Essentially, we need to assume that the problem in front of the panel is an "easy problem". In difficult scientific areas relevant to policymaking—such as climate change!—this may be an uncomfortable assumption. So solving one problem for applying the Jury Theorem has made another. Even in easy cases, we will face difficulties in applying this reasoning to expert panels as the numbers involved are usually small, so that the limiting behaviour that gives the theorem its force is not in operation.

3.2.1.5 *Error minimisation*

The simplest reason to use an average, when faced with a collection of predictions, has to do with expected error. It was presented neatly in a recent paper by Rougier (2016). If one has to choose between using the average and using a randomly chosen member of the ensemble, then we can show that the average will perform weakly better on mean squared errors—a common way of measuring error, as well as a close cousin of the Brier score. This follows from a simple application of Jensen's Inequality.

Lemma. ***(Jensen's Inequality)*** *For a real convex function F, numbers x_i in its domain and positive weights a_i, $i = 1, \ldots, n$*

$$F\left(\frac{\sum_i a_i x_i}{\sum_i a_i}\right) \leq \frac{\sum_i a_i F(x_i)}{\sum_i a_i}.$$

An important instance is $F(\mathbb{E}[X]) \leq \mathbb{E}[F(X)]$, for $\mathbb{E}$ the expected value relative to some probability function defined over a space on which X is a random variable.

Theorem 1. *Consider a collection of experts E_i, $i = 1, \ldots, k$, each of which makes n predictions, labelled E_{ij}, $j = 1, \ldots, n$. Let $\overline{E}_j$ be the average over the ensemble for the jth output: $\overline{E}_j = \sum_i E_{ij}$. We will treat $\overline{E} = (\overline{E}_1, \ldots, \overline{E}_n)$*

as an expert, and call it the average expert. Let O_j be the observed value for prediction j—i.e., the true value, to which we compare. The mean squared error for the ith expert is $MSE_i = \frac{1}{n}\sum_j (E_{ij} - O_j)^2$. The average expert's mean squared error is $\overline{MSE} = \frac{1}{n}\sum_j (\overline{E}_j - O_j)^2$.

The average expert's error is never greater than the average of the errors of each ensemble member:

$$\overline{MSE} \leq \frac{1}{k}\sum_i MSE_i.$$

The proof is trivial, relying on the convexity of the mean squared error function and a single application of Jensen's Inequality (Rougier, 2016). (Indeed, the result will hold for any convex error function.)

The upshot of Theorem 1 is that we have a weak reason to use the linear average of the reports. Suppose we have a panel of disagreeing experts, and we want to select or construct an opinion that will minimise error (MSE or another convex measure of error). If we have no information to distinguish between the experts, and we consider using either a randomly chosen expert or the average opinion, Theorem 1 tells us that we would do better to use the average opinion.

This is a limited but positive motivation in favour of using linear pooling. It is positive in the sense that it establishes a benefit to using linear pooling as opposed to a randomly selected individual report. It is also quite general: the result relies only on Jensen's inequality, and all that assumes is the convexity of F, and that the weights are positive. It works only for linear pooling: this gives us the summation, and therefore the theorem will not hold for other kinds of pooling. It does so on the basis of the error measure being convex, but as popular measures of error *are* convex, this is no great limitation. Importantly the proof does *not* depend on the distribution of the reports or their errors—it is not an error cancellation or symmetry argument. It is limited because it establishes linear pooling's superiority *only* relative to a randomly chosen report. If we can do better than random selection we will no longer be guaranteed the linear pooling has any advantage.

3.2.1.6 *It works: empirical success*

Finally, we come to the reason many practitioners of expert elicitation use linear pooling: it works. Or so it seems. In the empirical literature on expert forecasting there is a widely held view that averaging over available forecasts increases the accuracy of the forecast. Pokempner and Bailey (1970) reported in their book on sales forecasting that it was already then a common and valuable practice in business. Its use has been supported by numerous studies across different fields; Clemen (1989) provides an annotated bibliography of 200 papers on forecast aggregation and summarises the evidence thus: "combining multiple forecasts leads to increased forecast accuracy. This has been the result whether the forecasts are judgmental or statistical, econometric or extrapolation. Furthermore, in many cases one can make dramatic performance improvements by simply averaging the forecasts." In a more recent review, Armstrong (2001) found that in 30 empirical comparisons, equally weighted combined forecasts reduced the errors of randomly chosen individual forecasts by, on average 12.5% and up to 24% . "Under ideal conditions, combined forecasts were sometimes more accurate than their most accurate components" (Armstrong, 2001, p. 417).

This empirical basis serves as strong prima facie justification for linear averaging, though it does not tell us why and when it works. Armstrong himself suggests that the effect is due to error cancellation and he notes that a common theoretical basis between forecasters results in their forecasts having positively correlated errors. He therefore advocates for aggregating forecasts produced using different methods. So once again we will need to pay close attention to the details of the case before adopting the linear pooling strategy.

3.2.2 *Axiomatic characterisation*

In this brief section I will shift gear and ask a slightly different question: supposing that we plan some opinion pooling procedure, why should we use linear pooling rather than one of the alternatives? It is here that I will engage with the characterisation results from the opinion pooling literature.

Pooling functions can be characterised axiomatically, in terms of ostensibly desirable properties for any aggregation procedure. We can categorise these into three groups: structural, rationality and agreement properties. An example of a structural property is Eventwise Independence (EI). When an aggregation rule has this property, the aggregate opinion on issue X depends only on X and not on any other issues.

A paradigm rationality property is External Bayesianity (EB). Individual agents are at their best, the thought goes, when they are Bayesian ideal agents: perfectly rational, and updating via conditioning. If we have n ideal agents, and we pool their opinions, then a pooling procedure which obeys External Bayesianity ensures that when they receive new information, then "from the outside" the group looks like a single Bayesian agent—i.e., we should get the same result if new information is passed to (all the) individual experts, and the new individual opinions are aggregated; or if that information is passed to the "group agent," which then forms its new aggregate opinion (see Genest and Zidek, 1986, p. 119 for discussion). A variant is Internal or Individualwise Bayesianity (IB), which demands that we get the same result if new information is passed to just one expert, and the new individual opinions aggregated; or if that information is passed to the "group agent", which then forms its new aggregate opinion (Dietrich, 2010b).

Agreement properties concern the special status of agreement between the opinions. A weak version is the Total Unanimity Property (TUP), which says that if all the opinions are the same function so that the profile is just n copies of some P, then the aggregate must be P too. A proposition-wise variant is Unanimity Preservation (UP), which says that if all experts agree on the probability of X, then the aggregate opinion should be identical to their common probability for X. A special case is Zero Preservation (ZP), which says that if all experts assign X zero probability, then the aggregate should do so too. Finally, we are often concerned with the dependence structure of probability functions, and Probabilistic Independence Preservation (PIP) says that, if all experts agree that X and Y are independent, so too should the aggregate.

It turns out that these properties conflict. If we demand that our aggregation rule is EI and ZP, then it must be linear (McConway, 1981). So, do we want these two properties? Dietrich and List offer a strong pragmatic argument in favour of EI: it "is easy to implement, because it permits the subdivision of a complex aggregation problem into multiple simpler ones, each focusing on a single event. Our climate panel can first consider the event that greenhouse gas concentrations exceed some critical threshold and aggregate individual probabilities for that event; then do the same for the second event; and so on" (2016, p. 525). This is very helpful. In the elegant formalism of opinion pooling it is standardly assumed that the objects we are aggregating are fully fledged probability functions. But in fact they are the probabilistic assessments of experts, which may not have the richness of actual probability functions: experts may not be able to tell us how the probability of X changes, conditional on Y or Z. Our pooling function will have to be operationalised in a procedure carried out by these individuals (perhaps under the guidance of a facilitator), and allowing them to proceed eventwise will significantly reduce the informational burden of the procedure.

ZP too seems quite normatively compelling: if every individual agrees some event is impossible (or equivalently, its complement is certain) then it seems natural that the pooled opinion should be that the event is impossible. While in general there will be reasons to doubt that if $P^i(X) = c \; \forall i$, for some $X \in X$ then $F(P^1, \ldots, P^n)(A) = c$, the extremal case seems more compelling.[3] Expert opinion aggregation is difficult enough without seeking reasons to doubt the things on which the group all agrees are impossible/certain.

These are two compelling reasons in favour of linear pooling. In the next section I will review reasons put forward in favour of geometric pooling. After rejecting them I will turn to the fact that neither linear nor geometric pooling can preserve probabilistic independence, a problem which threatens the entire pooling approach.

3.3 GEOMETRIC POOLING

I will now turn to geometric pooling. I have found less in the way of direct motivation for geometric pooling than linear and so most of my discussion will focus on the characterisation of geometric pooling.

One direct motivation that is available for geometric pooling is the same, flawed, argument from approximate Bayesianism that we saw above: that is, that one can choose a likelihood function such that the result of a supra-Bayesian update matches that of geometric pooling. This argument fails for the same reasons that the linear version failed.

[3] Briefly: we might doubt that Unanimity Preservation is a good idea because independent experts may have different reasons for assigning $P^i(X) = c$. Suppose that they all started with a prior of 0.5, and $c > 0.5$. Then if they have independent "boosting" evidence that gets them up to c, we may well expect that our credence should be above c—indeed we could reason in this way without knowing exactly what the independent evidence is. So insisting that the aggregate credence must equal c seems misguided. This logic doesn't go through for $c = 0$, as that credence is reserved for things that are known to be false or deemed impossible (setting aside weird cases with measure zero events).

3.3.1 *Axiomatic characterisation*

The main arguments in favour of geometric pooling come from its compatibility with the so-called rationality constraints introduced above. In particular geometric pooling functions are Externally Bayesian (EB) and unanimity preserving without being dictatorial (Genest, 1984b). On the other hand, linear pooling functions can only be EB if they are dictatorial (Genest, 1984a). (Multiplicative pooling, which I will discuss only briefly, is the only *Individualwise* Bayesian pooling function.)

This trades against the simplicity of eventwise independence, which geometric pooling does not satisfy. But, as none of the major pooling methods satisfy all of EB, EI and UP (Nau, 2002), we face a choice between simplicity and "rationality."

The scare quotes are because I do not consider EB to be a rationality constraint. At its core it is about *appearing to be* Bayesian, presumably in order to gain the instrumental advantages of Bayesianism (avoiding Dutch books) rather than because there is a premium on the label. But, if one takes seriously the notion that supra-Bayesianism is the normative gold standard for responding to the reports of experts, then we should surely be given pause by the fact that a supra-Bayesian *will not generally be externally Bayesian*.

To see this consider a case with an agent with probabilities P, a single expert P^1 and an evidential proposition E. As before, let F stand for a pooling procedure. Let P_E be the probability function resulting from conditioning on E. Above, I glossed External Bayesianity as follows: we should get the same result if new information is passed to (all the) individuals, and the new individual opinions are aggregated; or if that information is passed to the "group agent," which then forms its new aggregate opinion. We can thus formalise EB as the requirement that

$$F(P_E, P^1_E) = F(P, P^1)_E \tag{6}$$

This is a constraint that makes sense in conditions of informational symmetry, where each expert has and is going to receive the same information going forward (Dietrich and List, 2016b, pp. 530-31).

My claim is that if F is supra-Bayesianism, considered as a pooling method, then the condition will not hold. Let's consider a single proposition X, and as before P's prior is x and P^1's is x^1. First, consider the right-hand side, pooling and then learning. The pooling in this case is a supra-Bayesian update: the agent updates as discussed in previous chapter: $P \mapsto P_{\ulcorner x^1 \urcorner}$. They then learn the proposition E and update by conditioning. The result is $P_{\ulcorner x^1 \urcorner, E}$. (The expert also learns E and updates, but this is irrelevant at the moment.)

Second, let us consider the left-hand side: learning and then pooling. The agent and the expert each learn the proposition E, and update by conditioning incorporate this information. The results: P_E, P^1_E. Then, they share reports; but now the expert's report is not x^1 but rather some posterior x^1_E. The agent now updates by conditioning, resulting in $P_{E, \ulcorner x^1_E \urcorner}$. Do we have any reason to expect these two results to agree in general? Surely not. The order of the subscripts doesn't matter, as conditioning is commutative, but in this second case the agent *received a different report*—x^1_E rather than x^1.

It is *possible* for them to agree, but as before this will depend on contingent factors such as the agent's likelihood for the two reported credences. So for some agents who happen to have the geometric equivalents of the very peculiar likelihood functions discussed above, they will *happen to be* Externally Bayesian. But this is no part of their Bayesianism, it is mere happenstance. (The same argument applies to Individualwise Bayesianity, where only one of the agent or expert receives the new information.)

One case in which agreement is known to occur is when the agent and the expert are in a particularly constrained informational position, so that when the agent hears a credal report they can infer what the expert's evidence is and what their prior was. This is what the setup of the Aumann (1976) agreement theorem guarantees, and Baccelli and Stewart (ms) have shown that under these conditions geometric pooling and Bayesian conditioning agree (and geometric pooling is the unique pooling method to do so). This is an extremely interesting theoretical result, but in general we have no reason to suppose people to be in the Aumann setup. It requires common priors and a shared informational structure, such that reports are "identifiable" in the sense that hearing a report allows an agent to uniquely determine the evidence used to produce the report.

If supra-Bayesians aren't Externally (or Individualwise) Bayesian, why should the rest of us bother with it? Well, for one thing, I don't think supra-Bayesianism is a good standard. But there is also a second more direct motivation for wanting to be EB: avoiding manipulation. If a pooling procedure is not Externally/Individualwise Bayesian, then it matters when the experts receive their information. The aggregate opinion will be different if the information is received before or after pooling. Someone who has the power to strategically disclose relevant piece of information at a time of their choosing will therefore be able to manipulate the aggregate opinion. This is a legitimate concern, and there may well be circumstances where it is determinative.

It is not, however, a reason to worry about Bayesianity when considering expert elicitation for policymaking. Why? Because in actual expert panel cases, *only* the "learning then pooling" operation ever takes place. Let us consider how Ade from Case 1 would conduct a geometric pooling procedure in practice. First, he must elicit each expert's opinion on the matter at hand (e.g., the relevant sea-level rise prediction). Second, Ade must elicit whatever information is required for the pooling procedure (e.g., whatever he needs to set the weights in the pooling function—discussed in the next section). As we're considering EB, let us further suppose that he ensures that the experts are as close as possible to a position of information symmetry, on an ongoing basis. With all this in place, Ade can produce an aggregate opinion.

There are a number of complications that I will note briefly here. First, Ade plans to selectively replace certain of his credences with these aggregate answers, not to adopt a full aggregate credence function as his own. He therefore needs to consider the issues discussed in chapter 2 as well as any rationality considerations introduced here. Second, recall that geometric pooling is not eventwise independent. The pooled probability for a proposition X depends not only on the expert's probabilities for X but also on their probabilities for other propositions. We can see this in the definition of geometric pooling by noting that the normalisation constant involves summing probabilities for each elementary proposition in the algebra. If these probabilities are not elicited, geometric pooling cannot be carried out. So, if one wishes to use geometric pooling, one therefore needs to carry out the elicitation in a specific manner: setting up an agenda of questions that one is interested in, determining the maximally specific propositions, and performing the elicitation at the level of these "worlds." This may require experts to evaluate complex combinations of events, whose probabilities they struggle to assess.

Let us set these aside and assume that the elicitation is successful and Ade has at hand the pooled opinion of the expert panel. Now suppose that new information comes to light. Having initially asked about X, Ade now learns that Y, some unforeseen circumstance, is the case. Ade realises that Y may influence the probability of X. He therefore desires to know the probability of X given Y. In the theoretical discussion above, we considered to possible procedures which the axiom of External Bayesianity demands produce the same result. The first is *learning then pooling*, the second is *pooling then learning*.

Consider first learning then pooling. Ade ensures that all the experts learn that Y, and re-elicits their opinions on X. Like a good Bayesian, he assumes that when each expert learns this new information, they will update their opinion P^i to Q^i, where, $Q^i(X) = P^i(X|Y)$. He then pools them once more to form $F(Q^1, \ldots, Q^n)$.

Now consider pooling then learning. Ade already has in hand the prior pooled opinion, $F(P^1, \ldots, P^n)(\cdot)$. He knows Y and wants to update the pooled opinion, calculating $F(P^1, \ldots, P^n)(X|Y)$. This requires treating the pooled object not as a simple number, the average of the numbers received from the panel, but as a probability function. Let's denote the pooled prior by $\overline{P}(\cdot) := F(P^1, \ldots, P^n)(\cdot)$ to emphasise this.[4] According to Bayes' rule, he needs to know the values for $\overline{P}(Y|X)$ and $\overline{P}(Y)$—note the bar; these are aggregate opinions on the likelihood of Y and a prior for Y. So Ade needs to have asked, back at the initial elicitation step, for each expert's opinion on Y, and for their conditional opinion on Y given X. (Or, equivalently, for the probability of X and Y.)

In other words, Ade had to see Y coming and must have done so for *any* new proposition that he learns. Given our non-ideal approach, I think it is reasonable to claim that, in most cases, he won't have done this. Even experts can't be expected to foresee every eventuality and provide the relevant probabilities required for updating on come what may. We might do well every now and then by covering the most likely possible developments, but assuming that we will always have this information is clearly untenable.

So in practice, only *learning then pooling* is available to Ade. The manipulation scenario we are meant to avoid assumes that the pooling is a one-time event. Information arriving before this event is learned by

[4] If you prefer to think of functions as sets: Ade will need to have access to a much wider portion of the function set.

each expert, and later their opinions are pooled. Information arriving after this event is "learned by" the pooled probability function. But in reality this latter procedure is difficult, if not impossible. Instead, policymakers will be forced to inform the experts and conduct the pooling again. But if this is the case, we are not open to manipulation.[5]

The above motivates against worrying about EB and IB, which are in turn the major motivations for using geometric and multiplicative pooling respectively. Of the major candidates, that leaves us with linear pooling. This is bolstered by arguments in favour of both EI and ZP in cases like Ade's, but only if one is not overly concerned about preserving probabilistic (in)dependencies.

Pooling in a way that does respect causal dependencies is significantly more complex than any method discussed here. Bradley, Dietrich, and List (2014) describe a two-stage procedure in which, first, the experts' causal judgements are combined, and second, their probabilistic judgements on the consensus causal graph are pooled. The causal judgement combination is itself an aggregation procedure with many possible methods, which can be characterised axiomatically. Bradley, Dietrich, and List present an impossibility theorem that shows that no aggregation method satisfies various compelling axioms. Therefore, one needs to undertake a case-by-case analysis of the relative attractiveness of each axiom, to determine which to relax in order to identify a viable aggregation function.

[5] This argument clearly does not apply to other cases where we *do* have a full probability function after pooling. If one is in such a situation, perhaps if using pooling in a machine learning or prediction algorithm, then more care needs to be taken with the risk of manipulation.

This procedure is at once very complex (pushing against our pragmatic desideratum) and highly idealised. It assumes that experts can specify causal graphs and associated probability distributions. But in real cases involving climate panels, the experts' reports will be based on the results of complex simulation models. The models themselves encode causal relationships between physical science variables but also depend in complex ways on computational and statistical techniques, approximations and idealisations. Bradley, Dietrich, and List's procedure may simply be impossible for such scientists to reproduce in practice.

This leaves us in a difficult position. Linear pooling is a procedure that is attractively simple and preserves intuitively compelling consensus judgements, but it does not preserve potentially important (in)dependence judgements. Bradley, Dietrich, and List's causal aggregation procedure allows us to preserve judgements of *causal* (in)dependence— which are the ones we care most about—but at the cost of significant complexity, perhaps so much that it is not practically implementable for anything beyond the simplest agendas.

3.4 WEIGHTS

Nevertheless, pooling remains widely popular, not just among philosophers but in expert elicitation procedures used by policymakers. I will therefore continue my analysis of it, assuming from here on that we have settled on linear pooling following the above analysis. Now that we know the form of the aggregate opinion, $\overline{P}(X) = \sum_i P^i(X)w_i$, we must specify its parameters—the weights w_i. Various philosophical questions are raised by this procedure, and I will consider them here. (Note that the discussion in this section is independent of the choice of pooling function. Linear and geometric pooling both explicitly use weights and multiplicative pooling requires weighting the experts in order to determine the calibration function from their priors (Dietrich and List, 2016a, S.9).)

Weights are typically thought of as assignments of credibility or trust. The problem of this section is how to assess the relative strengths of a panel of experts.

One natural suggestion is to weigh all experts equally. After all, as Ade is no expert himself it might seem that he is in no position to compare their expertise. But note that he cannot truly escape such comparisons: the selection of the panel involves a comparison between putative experts. Some will inevitably be ruled out, and others in; the latter are judged to be better than the former. Now, perhaps this judgement is possible, while further rankings between the panel members is not. But this just puts a lot of pressure on the in/out decision itself, as equal weighting in situations where the panel is small makes the result sensitive to the inclusion of a single "bad" expert. As already mentioned, most panels will involve only on a small set of experts due to limited time and money so this problem is a serious one. Individual experts therefore have a large impact on the output; the inclusion of a single "bad" expert can have a significant impact on the quality of the elicitation.

This worry can be addressed in an equal weighting elicitation, though with difficulty. Robustness tests can be introduced in which single experts are excluded from the set and the result is recalculated. Cooke and Goossens describe such a process:

> *Experts/seed variables are removed from the data set one at a time and the "decision maker" [the aggregate] is recalculated, to account for the relative information loss to the original decision maker. If that loss is large, then results may not be replicated if another study were to be done using different experts and seed variables. Discrepancy analysis identifies items on which the uncertainty assessments of the experts differ most. These items should be reviewed to ascertain any avoidable causes of discrepancy (Cooke and Goossens, 2000, p. 305).*

Ideally, we want an output which is stable under substitution of experts. Practically speaking, however, one cannot expect to be able to continue the process until such a stable output is found. Indeed, Cooke and Goossens seem to regard robustness analysis as an indicator of potential sensitivity—a warning—rather than as a guide to seeking a stable result. And this is reasonable; in domains with wide expert disagreement, robust results may well be impossible to find.

Given these concerns, the intuitive simplicity of the equal weighting approach loses some of its attraction. If we can find a method for assessing and weighting experts that is accessible to Ade from his position as layperson, we should prefer the weighted approach as it deals naturally with these problems (by reducing the impact of the worst experts). And indeed, in the literature on expert elicitation it is common to find methods for doing just this (e.g., (Cooke, 1999), (Aspinall, 2010), and (Cooke, 2018) which contains many examples of such applications). The core idea is to use the experts' track records on predictions within this domain. These can be their real track records of past predictions, or can be generated using a test. In most conditions, there will not be a convenient, assessable and comparable set of past judgements for a given expert panel, and so testing is the usual approach.

To get a sense for how this works, let us consider the "structured expert judgement" method developed by Roger Cooke. Roughly speaking, this process—which is conducted by an "elicitation analyst"—works as follows. First, one works with experts in the relevant domain to design a test of skill. This begins with identifying test variables: quantities whose prediction is relevant to expertise in the target domain. These variables must be measurable, and we must be able to set test questions that we know the answers to while the experts need to work them out. Once the questionnaire is developed, all the experts are tested. Their performance on the test is taken as an indicator of their level of expertise—weights are assigned in proportion with test scores. Once the experts have been assessed, the actual elicitation takes place: each expert is asked to make the prediction of interest, their predictions are weighted in line with their test performance, and the weighted average is then used by the policymaker (Aspinall, 2010; Cooke, 1999).

Consider a weather forecasting example. An elicitation analyst works with a group of meteorologists to determine how weather forecasting works and what counts as a good test. Suppose that they determine that a good test variable is chance of rain,[6] and that this is determined on the basis of three other variables: pressure, temperature, and humidity in the days before the assessed date. A test is devised, consisting of a number of predictive tasks: forecasting the chance of rain on 10 days, given 10 sets of input variables. These represent historical cases, where the analyst knows the answers. Each expert makes their set of ten predictions, which are then compared against the binary historical data (rain/no rain) for each test day. The tests are scored, and the experts weighted according to their performance. The elicitation is then carried out for the target variable: chance of rain tomorrow. Each expert makes their forecast, and their weighted forecasts are averaged to produce the final result.

3.4.1 *Choosing a scoring rule*

The important question left out of the above is: how do we score the test? Suppose an expert states a 60% chance of rain this afternoon, and it does in fact rain. How good was this prediction? It may be surprising that there is no straightforward answer to this question. A plethora of different approaches—different "scoring rules"—exist for probabilistic predictions. The scores they output will be used to weigh experts, so the scores need to be good proxies for the credibility of the experts, or the degree to which the policymaker ought to

[6] I'm using "chance of rain" in its common, idiomatic sense. I'm not interested in the question of whether meteorologists provide true chances or mere subjective probabilities, and this distinction shouldn't matter for the present discussion.

trust them. The different rules correspond to different ways of assessing the predictions the experts make, and therefore care is needed in selecting one which tracks what we think is important. The variety of rules and the technical differences between them makes this a challenging task.

Indeed, there are infinitely many scoring rules for probabilistic predictions and this variation is not merely a theoretical fact. Following a conference on forecast verification, the Australian Bureau of Meteorology (2017) compiled a list of 50 rules and techniques used for scoring weather forecasts. This list is broken down into seven categories of forecast (binary, multi-category, continuous, probabilistic, spatial, ensemble, and rare), and covers straightforward scoring rules as well as visualisation techniques and analytical approaches to measure forecasting success. Some are appropriate only for a single category, others have broader application. Choosing which to use, they suggest, is a nuanced and complex task.

In the field of expert elicitation (the origin of scoring rules, according to Cooke, 1999, p. 121) two families of rules seem most popular: quadratic and logarithmic (Douven, 2018). The Brier Score, the most famous instance of a quadratic rule, was developed in meteorology (Brier, 1950; Murphy and Epstein, 1967). The logarithmic score was proposed and defended by I. J. Good (1952), on the basis that it does not depend on the predicted values for any but the actual outcome—the Brier score, by contrast, depends on one's predictions for all outcomes.

To navigate the landscape of scoring rules, philosophers and statisticians have identified a number of desirable properties for scoring rules (the literature on this topic is very large, but see for example Savage, 1971, Cooke, 1999, Pettigrew, 2016).

- *Accuracy*: it is better to be closer to the truth. If the event occurs, it is better to have assigned it a higher probability than a low, and vice versa.
- *Calibration*: expert opinions should be rewarded for matching the data. e.g., It is desirable that a proposition X assessed to have probability $x\%$ be true $x\%$ of the time, out of the known observations of X.
- *Low entropy*: This is a measure of how "spread out" probabilistic judgements are. Lower entropy is favoured, as it is more committal about what will happen and therefore supports clearer belief formation and decision-making.

Accuracy and calibration are desirable for epistemic reasons; the first concerns truth-tracking and the second concerns responsiveness to evidence. Entropy is harder to categorise; it could be considered epistemic in the sense that entropy is measured relative to a set of data (cf. the notion of a maximum-entropy distribution for a given dataset). But the value of low-entropy predictions is pragmatic: they exclude more options, and thus give more direction on what to believe/how to act. In any case, it is of great utility to policymakers. These properties will trade off against one another given the uncertainty that all experts face when making predictions, e.g., they can attempt to make more specific predictions (lowering entropy), but at a risk to calibration and accuracy.

A further desirable property concerns the behaviour that the rule induces in experts.

- *Propriety*: Experts receive maximal scores iff they state their true opinions. There should be no benefit (from their perspective, given what they know) to submitting an altered prediction.

Propriety is a form of strategy-proofness; it ensures that there is no gain to be had by stating any opinion other than your true one in an elicitation process. This is important for iterative procedures, where the experts know how their opinions are being combined. If an improper rule is used, they can "game" the system by stating an opinion other than their true one, in order to pull the average closer to their true position (Cooke, 1999, Ch. 8 and 9). As our policymakers want to know the experts' true opinions, this seems obviously desirable.

I now want to highlight a sociological fact: there is significant disagreement over which scoring rule is best, either simpliciter or for a given situation. Different rules will often rank expert reports quite differently, leading to different pooled opinions. This makes the dispute over them a matter of real importance to the policymaker who has initiated the expert elicitation. But this policymaker is in a difficult position. The good-making features given above for rules are technical, difficult-to-understand properties and it is not clear how they should be traded off. Philosophers and statisticians, who are experts on these matters, disagree. Worse still, even for a given valuation of these good-making features, well-informed experts may recommend more than one rule as suitable.

This is bad news. Recall Ade, who began by facing expert disagreement in the domain of climate science. Having determined that he would aggregate the opinions of his expert panel, he has selected an aggregation method (linear averaging) and determined how to conduct an expert elicitation. However, when it comes to scoring the elicitation test, he has come up against yet another expert disagreement, this time amongst statisticians and philosophers. How will Ade resolve *this* disagreement? Surely not by aggregating, which would lead to a regress. But any *non-aggregative* method of choosing a scoring rule requires deciding between disagreeing experts. So, for whatever method one might suggest for resolving this second disagreement, we may well ask why we do not simply apply the same considerations to the "first-order" problem of expert disagreement, thereby bypassing the pooling procedure entirely.

Despair would be a mistake, however. This same second disagreement (over scoring rules) emerges for *all other* expert disagreements, supposing that we decide to resolve them using weighted opinion pooling. Nothing about the situation above that led us to this scoring rule disagreement depended on the *domain* of the expert disagreement—it will crop up whenever we aggregate.

This is good news! The problem of expert disagreement is at heart one of epistemic resources: laypeople cannot afford to acquire the skills needed to decide for themselves in all the myriad expert domains. Socially, we rely on a distribution of cognitive labour such that there are different domains of inquiry, each with its experts, to whom the rest of us (more or less) defer. The problem is that when these experts disagree, we are at an impasse. But, if aggregation is a viable method for solving expert disagreement, then each of these distinct instances of expert disagreement lead to the same second-order problem of disagreement between statisticians and philosophers over the correct scoring rule. So we have identified a specific domain of inquiry (viz. scoring rules) which is of great utility in resolving the social epistemic problem of expert disagreement.

When assessing how to invest their limited cognitive resources, policymakers have a strong reason to invest in acquiring *this* expertise, of expert elicitation and in particular scoring-rule-choice. It will allow the policymaker to tackle a great many individual cases of expert disagreement, in distinct areas of expertise.

For laypeople this is an unrealistic recommendation. But policymakers like Ade occupy a specific social role: making decisions on behalf of others, using the best scientific evidence. It does not seem unreasonable to me that this form of (meta-)expertise be considered part of that job. Now, as discussed above, there are expert elicitation practitioners who can conduct these elicitations for him. But there are multiple practitioners, each promoting their own version of the service. For example, Roger Cooke has defended his method in some detail, including a particular approach to expert scoring (using a chi-squared test); while others use different rules. Without gaining expertise himself, Ade faces the regress above.

3.4.2 *The role of values in choosing a rule*

There is another reason why Ade ought to participate in the choice of scoring rule directly: in addition to their epistemic properties, scoring rules are differentiated by non-epistemic values, and it is part of Ade's role as policymaker to supply those values here.

To establish this, let us begin with a quick review of the notion of "inductive risk" and the role of non-epistemic values in probabilistic inference. We start with a simple case of test design. Suppose we wish to design a pregnancy test. We know that any mechanism used to detect pregnancy will have some error involved. So we must decide: if our test detects pregnancy successfully $x\%$ of the time, how high should x be for the test to be a good one? The answer, as is well known, depends on what is at stake. Consider the standard error table below. When we design the test, we need to consider the two possible kinds of errors, false positives (test reads positive, patient not pregnant) and false negatives (test reads negative, patient is pregnant). A trade-off between these errors depends on their impacts on the people involved. The (dis)utilities of these outcomes guides our test design: we attempt to maximise expected utility.

Table 5: Standard error table

	Observed positive	Observed negative
Forecast positive	True Positive	False Positive
Forecast negative	False Negative	True Negative

The important fact is that this is not an epistemic question. The values used do not concern truth-tracking, or responsiveness to evidence, or consistency. Instead, we reflect on the disutility of being pregnant but not knowing it, or of the anxiety caused by a false alarm.

This basic thought underlies a classic argument for the ineliminable role of values in science (Churchman, 1948; Rudner, 1953). In the ordinary course of science, the argument goes, scientists accept or reject hypotheses. This involves a decision about how much evidence is required to make either judgement. This level of sufficiency represents a trade-off between possible error types: require too much evidence, and one risks rejecting true hypotheses; but require too little and one risks accepting false hypotheses. And so, it is argued, the practice of science similarly depends on non-epistemic values. These determine the impact of the different kinds of error: "How sure we need to be before we accept a hypothesis will depend on how serious a mistake would be" (Rudner, 1953, p. 2).

There are different versions of this argument. In some, the contention is that there is *no way* to make a statistical inference without making a non-epistemic value judgement. Others argue that while one might make this judgement without explicit reference to non-epistemic values (typically, by convention or ignorance), to do so would be *irresponsible*: in the choice between arbitrariness and value-ladenness, the responsible scientists ought to think carefully about the values underpinning their research.

It is not required for my point here that this argument succeeds for science in general. What we need is to see how it applies to scoring rules. Here, we are assessing predictions themselves rather than using predictions to make inferences. We have already seen that we can assess them on the basis of certain epistemic properties such as calibration and entropy. But we can also assess them in terms of the kinds of errors they produce on the test data.

Let us stick with the simple case of binary forecasts to begin. For a set of predictions, we can track the standard error types shown in Table 5. The resulting statistics on each prediction's error performance can themselves be used as scoring rules (Australian Bureau of Meteorology, 2017). For example, consider the following:

- *"Accuracy":*[7] *(True Pos + True Neg)/Total*. This measures the proportion of correct predictions made, relative to the total number made. It has the virtue of being simple, but it fails to differentiate between True Positives and True Negatives. In many cases, one of these will be much easier to achieve than the other. Consider a binary prediction of whether London will be hit by a snow storm. Predicting "no" will be much more likely to succeed, and thus a high "accuracy" score will be easy to achieve. What we would want in such a predictor, however, would be sensitivity to True Positives in particular, which are much harder to come by.

- *Sensitivity: True Pos/Observed Pos.* This measures the "hit rate" of the prediction, the proportion of positive instances that it correctly predicted. This would be more suitable for the snowstorm case than the above. But here, we ignore False Posi-

[7] This standard term for this statistic is somewhat confusing in a context where "accuracy" also refers to a more general property but as the statistic will not occur again in this text the conventional name is used here.

tives. So in cases where *those* are the important errors, this score would be misleading.

And so on. For binary tests, the Australian Bureau of Meteorology (2017) lists 12 test statistics that can be used as scoring rules, individually or in certain combinations. If we suppose for the moment that everything else is equal between these rules then, as before, the decision about which we should use will depend again on the stakes, i.e., what we care about. (Simple test statistics don't make for very good scoring rules on other grounds, so the purpose of the example is merely to illustrate how scoring rules differ in terms of values.)

This point generalises to scoring rules for probabilistic predictions. This should be no surprise, for in his original article on values in science Rudner notes that even if the role of the scientist is only to provide probabilities (rather than accept hypotheses), the value question remains. For then, the problem merely moves back a step to "the acceptance by the scientist of the hypothesis that the degree of confidence is p... [and the] acceptance of hypotheses does require value decisions" (Rudner, 1953, p. 2). Epistemologists have recently recognised this: Moss (2011) makes this point in a discussion of peer disagreement, and in a recent paper Babic (2019) develops a theory of "epistemic risk" in the context of accuracy-first epistemology. A brief discussion of Babic's theory will serve to illustrate how Rudner's point about accepting a probability function transfers over to rules for scoring such functions, in contexts where the goal is to combine those functions.

Babic treats a scoring rule as a two-place function, denoted $s_v(P(X))$, which measures the inaccuracy of the probability assignment $P(X)$ when the outcome is v, where $v = 1$ if X is true and 0 otherwise. In line with our epistemic value of accuracy (in the sense of the last section, not the test statistic just mentioned), s_1 is an increasing function of P, and s_0 is a decreasing function. These functions are assumed to be continuous, and as they are both defined on the domain $[0, 1]$ the intermediate value theorem tells us they will intersect at some P^*, for which $s_1(P^*) = s_0(P^*)$. This is an assignment of probability that guarantees a certain inaccuracy score no matter the outcome, and in that sense is "risk-free". The "risks" here are risks of errors: assigning a high probability when the proposition turns out false and vice versa.

Where the risk-free point occurs depends on the nature of the rule s_v. Babic discusses a class of symmetric rules: consider two probability assignments to X, $P(X)$ and $P'(X) = 1 - P(X)$. Then s_v is 0/1 symmetric iff $s_1(P(X)) = s_0(P'(X))$. So for example if one forecaster says the chance of rain is 0.4, and another 0.6, a 0/1 symmetric score will assign them the same score no matter the outcome. For such scores, the risk-free point is $P^*(X) = 0.5$. For rules that are not symmetric the risk-free assignment is different, reflecting their different valuations of the two ways of being wrong. None of the epistemic values that I introduced above, nor any that Babic considers, force us to use a symmetric rule.

The value of Babic's work is that he clearly illustrates that something we might take for granted, or do unthinkingly, is just one possibility among many. “It is not enough, therefore, to declare that we should seek truth and avoid error. Such an epistemic norm is underspecified. We need to decide further how to trade-off the potential costs of different types of mistakes” (Babic, 2019, p. 19). While popular scores such as the Brier score are symmetric, this is merely one way of achieving the “epistemic” goals of scoring. So either we need an epistemic argument for symmetry, or we must accept that the choice of risk-free point is non-epistemic.

One defence of symmetry, by Pettigrew (2016), has already been discussed in section 2.7, where I expressed my reservations about it. I do not know of any others. As I see it, the dialectic usually involves philosophers saying that non-symmetric rules involve pragmatic considerations, which they want to eschew. Therefore, they say, we must use a symmetric rule. But this begs the question, by simply assuming that something about symmetry is non-pragmatic. Symmetry values the two risks of error equally, but I see no reason to accept that this is a neutral position. It is a substantive value judgement, and one that may not always be appropriate.

Any choice of rule, I think, involves an implicit value commitment. Given that this is inevitable, we should ask: whose values should they be? In our case, it seems clear that the policymaker ought to supply them. These should not be *their* values, as an individual; the policymaker acts as social planner, deciding with an eye to social welfare. After all, the policymaker wants to form the aggregate opinion in order to make policy decisions. In so doing, they will use these probabilities alongside social welfare considerations. The problem is: different options for aggregations exist. Bracketing disagreements over epistemic values, they differ in how they treat the different kinds of epistemic risks. The claim here is then simply that the policymaker can reasonably select the aggregation that best caters to the risks they care about qua social planner.

This may feel uncomfortable. After all, Ade undertook this expert elicitation in order to learn what will happen. Now, he has found himself crafting the view on what will happen according to his values. Are Ade's desires influencing his beliefs in an improper way? I think not. Ade is using values here in what Douglas calls an "indirect role" in science. A direct role is when values themselves act as a reason to, for instance, accept a claim. They act as, or supplant, evidence. In the indirect role, values only "determine the importance of the inductive gaps left by the evidence" (Douglas, 2009, p. 96). As evidence increases and uncertainty decreases, the scope for values to play an indirect role naturally decreases too. The indirect role, Douglas argues, poses no threat to the integrity of science and indeed is indispensable to it. So, what Ade does here is, on Douglas's view, no different from the healthy functioning of science. The uncertainty he faces is due to the disagreement between the experts. Should that disagreement narrow, there will be less scope for value-judgements to play a role in determining the aggregate opinion.

We have seen that a great many expert disagreement problems can be reduced to one: the selection of a scoring rule, to be used to evaluate the relative expertise of the panel in aid of an aggregation procedure; said aggregation to use linear averaging. Expert disagreement is not, it turns out, a purely epistemic problem. This suggests two things. First, it is valuable to invest in the expertise of scoring rule selection. For policymakers like Ade, the investment of cognitive resources here is far more valuable than investing in any particular domain (hurricanes, earthquakes, forest fires) which he may encounter in his role. Second, there is valuable work to be done in explicating the value-commitments of different scoring rules. Not all parts of the choice between them is value-based; the nature of the problem (e.g., the rarity of the events in question) plays a role, as does the blend of epistemic values deemed appropriate. Where these characteristics do not constrain the set of scoring rules to a singleton, there is no avoiding value judgements. But we do much better to surface them, and deliberately employ the values of the policymaker, than to allow them to be made implicitly by statisticians.

There is a more fundamental concern, however. The structure of the aggregation procedure is pushing us toward this interaction between our values and our epistemic attitude. The reason that we choose a single set of values (the policymaker's or someone else's) is that we need them in order to select a scoring rule, which we need in turn in order to conduct opinion pooling. But if the epistemic values under-determine the choice of scoring rule, and we wish to maintain a strict separation between epistemic and non-epistemic values, then an alternative response is simply to reject opinion pooling in favour of an approach that does not require us to collapse the set of expert judgements down to one. In chapter 5, I therefore turn to a decision theoretic treatment of expert disagreement. There, the profile of expert reports is taken as an input to an imprecise decision theory in which there is no requirement this disagreement to be resolved before the policymaker's decision-making begins.

Before that, I want to discuss an important case of expert disagreement that will play a key role in a case study later in this thesis. That case is when the "opinions" that disagree are the outputs of scientific models, themselves constructed by scientific experts whose uncertainties and disagreements are partially represented by their models.

4

SCIENTIFIC MODEL ENSEMBLES

4.1 INTRODUCTION

This chapter is the transition point in this thesis, between the first part that is largely epistemology, and the second which involves more decision theory. Let me therefore recapitulate the central problem, to set up this chapter's discussion. I am interested in decisions made by policymakers, who consult with experts. The basic decision-theoretic setup is familiar. Decision-makers face a set of options: things they might do. Which of these they ought to do depends on how the world might be, and what the consequences of each option would be, given various states of the world. I am considering situations in which the decision-maker, a policymaker, consults with scientists to get information on how the world might be. The scientists provide them with information to fill out their state-partition, and to provide them with (information to fix their) probabilities over those states.

When policymakers, or their scientific advisors, consult with scientific experts they often discover significant uncertainty. In the previous chapters, I looked at cases where this uncertainty manifests as disagreement between experts. In this chapter, I turn to another common manifestation: uncertainty in and surrounding scientific models.

Models are key tools for scientists in many fields and are often the vehicle by which the advice for policymakers is generated. I will here discuss how models interact with and represent scientific uncertainty, and how they are used for decision support.

I begin with a brief introduction to scientific models and in particular to collections of models, or "ensembles." I then introduce my case study, which involves models of hurricanes in the North Atlantic. Using this case, I discuss averaging model outputs and develop some particular concerns for the use of opinion pooling tools (introduced in chapter 3) in cases involving models. I then draw specific conclusions for the case of the hurricane ensemble. This lays the ground for the next chapter, which introduces a decision theory and applies it to a case of model disagreement.

4.2 ENSEMBLES OF SIMULATION MODELS

In chapter 7, I present a general overview of the philosophy of scientific modelling and so in this introductory section I will focus on one particular class of models involved in supporting scientific policymaking: simulation models of complex systems.

Let us begin by making more precise the notion of a simulation model. Wendy Parker provides a clear starting point: "a simulation [is] a time-ordered sequence of states that serves as a representation of some other time-ordered sequence of states; at each point in the former sequence, the simulating system's having certain properties represents the target system's having certain properties" (Parker, 2009b,

p. 486). As this is somewhat abstract, I will elaborate. A "state" is a total description of some system's properties. Parker's definition speaks of two such systems; typically one will be a physical system in the world (called the target), and the other an abstract system (the model). Simulation models are thus instances of theoretical models (as defined in chapter 7): a simulation is a "theoretical model materialised in a computer," consisting of numerical equations and an interpretation of the mathematics (Petersen, 2012, p. 8).

Models are often thought of in terms similar to tools, or artefacts, rather than as true descriptions or accurate representations of their targets. Modellers strive to make them useful for certain purposes. These are often quite specific, e.g., predicting rainfall in a particular region, or predicting hurricane damage risk for a part of the US's eastern coastline. Their success criteria are therefore not taken to be truth or accuracy, but adequacy for their intended purpose (Parker, 2009a; Teller, 2001; Weisberg, 2007b) .

Theoretical models (including complex simulation models) have a mathematical representation, in which (some of) the system's properties are represented by variables in algebraic expressions. As described in chapter 7, this allows us to speak of the model as having various properties typically associated with physical systems (mass, speed, etc.). Not all properties of a model are relevant, however; models come with a "key" (implicit or explicit) stipulating which properties are relevant, and what features of the target system they map onto (Frigg and Nguyen, 2016).

Given this, we can think of complex simulation models as having the following parts. First is a structural-dynamical description of the system; this is a description of how the target system works, how it evolves from one state to another. In many scientific domains these dynamical relations are captured with differential equations, involving the variables mentioned above as well as parameters (which are fixed quantities, supplied exogenously). As simulation models are realised as programs on digital computers, these equations need to be discretised and implemented in computer code. When I refer to a model's structural-dynamical properties, I am referring to all of the foregoing.

The structural-dynamical part of the model captures those parts of the system's dynamics that the modeller (a) has a theoretical understanding of, and (b) regards as important to their current purposes. Other parts of the system might be "parametrized"—that is, represented in the model not via (what the modeller takes to be) an appropriate dynamical description, but in a simplified form. These simplified representations often take the form of a single parameter or set of parameters—hence the name—whose values are supplied by measurements, simple calculations, or other models. As mentioned above, dynamical equations often involve (other) parameters too. Thus the second general feature of simulation models is a set of parameters.

Third, the simulation requires a set of initial conditions, specifying the state of the system at the time the model is initiated. The structural-dynamical description, parameters, and initial conditions are sufficient information to "run" the model; that is, to perform the calculations required to specify states following the initial state in the time-ordered sequence. The model then generates a set of outputs, the state or states the model evolves into given the foregoing. [1]

In sciences dealing with complex systems, it is common to encounter a range of different models representing the same system. Such models might disagree deeply, over the structural relations in the system; or in shallower ways, over the values of parameters or initial conditions. Prominent examples are the CMIP5 ensemble of global climate models and, as we shall see, ensembles of hurricane models for the North Atlantic. In some cases, model ensembles indicate disagreements amongst scientists; in others, they reflect variation in acceptable methods of model construction. In either case the ensemble represents (at least partially) scientific uncertainty about the target system.

At the most basic level, one might have multiple models of the same system which describe it using different structural-dynamical descriptions. These are called multi-model ensembles (MMEs). MMEs may arise because each model was built for a slightly different purpose, but is capable of answering questions about the same decision variables. On the other hand, an MME may arise because each model was built to answer the *same* question(s) about a set of decision variables, but there are disagreements between scientists about the theory or modelling technique. Few sciences regard themselves as complete, and so scientists regularly have reason to doubt that they

[1] The above is sometimes taken to be a *general* picture of models in science, such that one can define a model as a set of trajectories in a state-space, determined by the dynamical equations discussed above (e.g., van Fraassen (1980), Lloyd (1984)). I make no such general claim, and take myself to merely be providing an introduction to the kind of models I discuss in this chapter.

have access to the "true" structural-dynamical features of the world. This leads to the consideration of a variety of models, with distinct structural-dynamical descriptions (and, therefore, often also different parametrizations and parameters).

Uncertainty about parameters gives rise to a different kind of ensemble. As mentioned above, "parameters" refers to constants in algebraic expressions, playing two different roles in models. The first kind mark a boundary in a scientific theory; they supply numbers which must be experimentally provided in order to make the theory complete. For example, the masses of the leptons and quarks in the Standard Model are measured in experiments and supplied to the theory, rather than being predictions of it. The second role is as representation of science left out of the model. Models often operate at a particular scale, and interactions below this scale will be represented in the model by a parameter: a number which summarises the lower-level interactions. For example, cloud dynamics are typically represented in climate models by a parameter, rather than being modelled in full.[2]

There is uncertainty about each of these kinds of parameters. For theoretical boundary parameters, there are the typical sources of uncertainty in measured values. For modelling simplification parameters, this measurement uncertainty will still be present and will be compounded by uncertainty in the approximation technique used to

[2] In climate science, people speak of "parametrisations." "A parametrisation is a mathematical model that calculates the net effects of these 'unresolved' processes on the processes that *are* directly calculated in the forecast model (the 'resolved' processes)," (Petersen, 2012, pp. 25-26). They transform a whole dynamical process into a parameter (number, or simple function) that replaces it as an approximation in the overall model—as described above.

generate the parameter. There will also be error introduced by using a single, typically static, value in place of the full low-level dynamics. We can explore and quantify parameter uncertainty with a sensitivity analysis, also known in climate-modelling as a perturbed physics ensemble (PPE). This takes a single model structure and varies the parameter values to generate a range of instantiations of this model, each with a different outcome.

Finally, there is uncertainty in the correct initial conditions. This is due to measurement uncertainty and will be compounded by modelling decisions about what to leave out. Modellers typically aim to capture the most important contributions to system dynamics, and to exclude small-size effects in order to reduce the complexity of the model. Scientists often explore this uncertainty through an initial conditions ensemble (ICE). For example, the ensemble forecasting technique employed by the UK Met Office is an ICE (Met Office, 2017).

So in summary we have:

- Multi-model ensembles, which collect "genuinely different" models, each representing a different view on which structural relations hold of the world.

- Perturbed-physics ensembles, which collect models which agree (for the most part) on the structural relations but disagree on the values of important parameters.

- Initial conditions ensembles, which collect instances of "the same model" which are initiated using different data, describing different initial conditions of the modelled system.

In what follows I will focus on multi-model ensembles. This is for two reasons. First, multi-model ensemble studies play an important role in climate science and its intersection with policy. The IPCC uses multi-model ensembles in its assessment of climate change, and hurricane modellers use multi-model ensembles when providing regulators and insurers with medium range forecasts of the probability of hurricane damage. Second, and relatedly, the distribution of outputs of the members of a multi-model ensemble tends to have a special status in multi-model studies that it does not have in perturbed-physics or initial-condition studies. Agreement between or clustering of model outputs is often taken as a symbol of the robustness of those results, when occurring in MMEs (Lloyd, 2015; Parker, 2011).

I will primarily consider cases in which models provide probabilities for events, or probability distributions over variables of interest. In the hurricane case study that I will later consider, this is one important output from the models. In other cases, such as global circulation models for climate change (GCMs), the models themselves produce values for variables of interest, and ensemble methods are used to produce probability distributions for those variables. (This latter use may not be well justified, as we will see later in this chapter.)

4.3 DECISION-SUPPORT WITH MODEL ENSEMBLES

Ensembles enter decision-making because decision-makers seeking expert input have no principled way of reducing the uncertainty and disagreement that ensembles represent. Decision-makers, who are typically non-experts, cannot select a single modeller, or model, or set of parameter values, or initial condition. They are in a worse epistemic position than the experts who constructed these models, and whose uncertainties and disagreements gave rise to them. In some cases, like GCMs, those experts regard the multiple models as having roughly equal prima facie plausibility as tools for the relevant task, such as projecting future climate change (Parker, 2006, 2011). In the face of significant uncertainty, scientists and advisors therefore require tools to provide useful decision-support starting from a model ensemble.

What makes decision inputs "useful"? In my introductory remarks above I described scientists as providing information about the relevant possibilities (often called states or events) and the probabilities of those states. This way of framing things comes from the standard decision-theoretic conception of a decision problem, in which uncertainty about the world is completely captured by a single probability measure over the relevant states (often in the form of a probability distribution over a variable of interest, like sea-level rise).

In certain decision-theoretic presentations, the probability distribution represents the agent's partial beliefs about the relevant states. It summarises whatever it is they believe, and so the uncertainty it represents is *their* uncertainty. In more practical applications, where decision theory is put to use in assisting various people with decisions they face, the probability distribution is more often taken as an input to the theory that is constructed using the best evidence available. This may involve significant subjective judgement (perhaps on the part of the agent, perhaps by their advisors), but it isn't a representation of the beliefs of any individual.

This latter circumstance corresponds more closely to my interests, as I regard the Bayesian orthodoxy to be impossible for any real agents particularly when faced with expert disagreement (as discussed in section 2.2). So I am interested in how real decision-makers arrive at the kinds of inputs required by decision theory and (in this chapter) in particular when faced with an ensemble of models, each purporting to generate the relevant probability measure.

However this is done, we can make some general comments about desirable features for a probability function to have, given that it plays the role it does in our decision theory. On the one hand, we would like to minimise uncertainty. Ideally we would want certainty about which state will occur, and failing that it would be helpful to have a state, or small set of states, that is clearly most likely. (Graphically, this corresponds to a narrow distribution curve with a high peak.) This is helpful to decision-makers as it gives them a more definitive sense of what is likely to happen, and therefore of which option is best. On the other hand, as a probability distribution becomes more and more peaked we might worry that it is likely to get things wrong by over-committing—assigning too much probability mass to a narrow set of options and neglecting others which are possible. Such probability measures can lead us to courses of action that do not perform well if things turn out differently than how we thought. We can summarise this by saying that a representation of uncertainty should be "robust," in the sense that small variations from our assumptions do not lead to results that are drastically wrong.

As decision-makers make use of the results from a model ensemble, they will need to navigate the trade-off between these two considerations. Given a relevant collection of models, one simple way to derive a single probability function is to select a single model for use (presumably, one assessed to be best for the purpose at hand). The obvious problem with this method is that the decision-maker has little basis for the choice and therefore is likely to get it wrong. Averaging the outputs of the models is another way; one which is often presented as more robust than selecting a single model. I will argue that this is not the case for an important class of models.

Indeed, I see the problem as being the demand by orthodox decision theory that uncertainty be represented by a *single* probability function, and so in the chapter after this one I will move away from this requirement. Recent decades have seen the development of numerous decision rules for situations in which decision-makers face what is known as "ambiguity", when precise probabilistic estimates of all decision-relevant quantities are unavailable (Gilboa and Marinacci (2013) and Heal and Milner (2014) provide surveys). There is also a nascent decision-theoretic literature on model uncertainty (see in particular Marinacci, 2015). In chapter 5 I will turn to the development of a decision-theoretic approach, called the "confidence" approach, for cases where the probabilistic information available comes from a model ensemble.

4.4 WHY AVERAGE?

As before, I will begin with a consideration of why it is that one would think to average the results of multiple models. Many of the reasons for averaging are the same for models as for opinions. The relevant reasons to average from chapter 3 are: sampling analogies, other convergence arguments, error minimisation, and empirical success. These all run into problems, as I will briefly show—either the same problems as for opinions, or particular model-specific ones.

Sampling. How would the sampling motivation go for models? One option is that model results are like measurements and thus the collection of those results in like a sample. Another option is that our collection is a sample of models, either the models that actually exist or those which might. Model ensembles are no more like samples than profiles of expert opinions are, however. As discussed above, model ensembles are either deliberately constructed or formed by collecting those few models that exist for a system. Neither case constitutes random sampling—in either of the senses just described (Tebaldi and Knutti, 2007).

Once again, it is also unclear what the population we are drawing from is, or why we want to identify the population mean. Models are representations of a system; some are faithful in certain respects; some are useful for our purposes. The range of "all models of a system" seems boundless, and I don't have an intuition that its mean is "the true model," whatever that is. Suppose that we can find a way of specifying a set of fairly accurate, useful models. In what sense is the small set of models we happen to have a "random sample" of this population? There is no reason to believe model generation—a process carried out by scientists who know one another and work in a particular disciplinary matrix—will meet the technical definition for a random sample: a random sample is one in which every element of the population has a non-zero probability of being selected as a member of the sample, according to a probability measure on the population that is either known or can be determined. It is implausible that the relevant scientists are equally likely to generate each of all the plausible models, or that we could construct a distribution describing their probabilities of "selecting" particular models. (In this discussion, I agree broadly with Murphy et al., 2007 and Parker, 2010, 2018, though the nuances of my argument are slightly different.)

Convergence. How might we apply the Condorcet Jury Theorem to model ensembles? Recall that it involves relying on the majority position of a set of voters, which can be shown to converge on the true position given some assumptions. Those assumptions include that the voters are $> 50\%$ likely to vote correctly, and that their votes are independent. (There were significant further technicalities in the discussion above that I won't rehash in detail.) An application of the CJT to models would put model results in the place of votes, and think of the average as a kind of majority position. Or alternatively, we could put it to a different use and could consider model results as votes, and evaluate whether responses lie above or below some threshold.

The problem here is that if model results are in the position of votes, then models must be independent. But, as alluded to above, it is implausible that models will obey any of the independence conditions that are required by the CJT or variants of it. Modellers share training and experience, they exchange portions of code and workarounds to common problems, and their models are based on common information (from the physical science basis to previous published results) (Parker, 2018, p. 283). Empirically, Parker (2018, p. 284) reports that "recent investigations have found that errors in simulations of past and present climate produced by today's state-of-the-art climate models show significant correlation." So we should not believe that the CJT or some variant can support model averaging.

Error minimisation. The simple error minimisation argument that I made for convex measures of error holds for models—indeed, Rougier (2016) develops it in the context of climate models. But recall that result's limitations: if we have no information to distinguish between models, and we consider using either a randomly chosen model or the average model, them theorem 1 tells us that we would do better to use the average model. I will return to this point in chapter 5; while I concede that this is *sometimes* a reason to use the average model, it is not a reason to use *only* that model (if robustness can be gained by using more than one output).

4.5 CASE STUDY: HURRICANE INSURANCE IN THE NORTH ATLANTIC

For people living in Florida, or on a Caribbean island, the risk of hurricane damage to their home is one of the most serious they face. Naturally, an insurance industry has grown around this risk, offering home-owners protection against the various forms of destruction hurricanes can bring. Residents buy insurance policies that guard them against such damage, frequently at high cost: a house valued at £120,000 will cost £2,500–6,000 per year to insure.[3]

The price is so high in part because hurricanes cause significant damage. But it is also because insurers are so uncertain about how risky hurricane damage is to insure. If you want to sell insurance against something the recipe is simple, with just three ingredients. First, you need the probability of the event you're covering (the hurricane). Second, you need an estimate of how damaging these events are when they occur—how much damage, in pounds, does the average hurricane cause to a £120,000 house? Third, you need to obey insurance regulations that tell you how much money you need to have available at any given time. These rules exist to ensure that insurance companies don't go bankrupt when catastrophes occur and have the money to pay for claims when customers make them.

But the first two ingredients are difficult to work out for hurricanes. Calculating the probability of destructive hurricanes requires a detailed understanding of the science of meteorology, as well as

[3] Estimate from https://www.sapling.com/7958883/average-cost-hurricane-insurance

complex statistical and numerical techniques. Estimating the vulnerability of a building to hurricane damage—in order to determine the monetary value of the damage—requires knowing how it was built, and how the building materials will withstand the wind and water effects of the storms. This typically requires knowledge of local building codes, estimates of compliance with those codes, and engineering studies of building vulnerability.

The scientific challenge of predicting hurricanes raises some surprising philosophical challenges. I explored these issues as part of a research collaboration with scientists working for a large UK-based insurance and reinsurance company. As part of its US property insurance business, the company offers cover for damage resulting from hurricanes.

It is often not economically efficient for insurers to invest in the expertise and capabilities required to price this insurance. Instead, they buy predictive models from commercial modelling companies. These companies employ teams of environmental scientists, statisticians, and programmers to construct simulation models to determine the probability of hurricane "landfalls" along the US's Atlantic coast.[4]

The modelling firms face a problem: there is significant uncertainty in hurricane modelling, derived in part from disagreements about the underlying science. The result is that there are multiple models representing the same system. The Florida Commission on Hurricane Loss Projection Methodology carried out an assessment of the modelling industry using an ensemble of 972 models (FCHLPM, 2007; Guin,

[4] In 2015, the Florida Commission on Hurricane Loss Projection Methodology received submissions for approval from four private firms: AIR, Applied Research Associates, CoreLogic, and Risk Management Solutions (FCHLPM, 2015).

2010)! Risk Management Solutions (RMS), a leading modelling firm, uses an ensemble of 13 models to generate the "Medium-Term Rate," their preferred prediction of hurricane landfall frequency (Sabbatelli and Waters, 2015).[5]

Any company selling models to insurers must decide how to navigate this landscape. Which model(s) should they build as part of their offering? Offering more than one model better represents the landscape, but presenting insurers with a collection of models creates a further problem for them: how does one decide when faced with not one model-probability but 13 or 972? The most common solution when working with ensembles is to average the outputs from each model.

To add specificity to the problem, and show how it arises in an important real-world application, I will now give a brief overview of the RMS model ensemble. RMS is a useful example because they are a leading hurricane modelling firm, and because they are open about their use of an ensemble of models: their Medium-Term Rate is the average of outputs from "13 individual forecast models, weighted according to the skill each demonstrates in predicting the historical time series of hurricane frequency" (Sabbatelli and Waters, 2015).

Let me start with an explanation of how such models work, in order to illustrate how the model ensemble arises. Natural catastrophe modelling is complex, and my treatment here considers only a small part of a typical "nat cat" model.

[5] The uncertainty is present at all timescales. For a discussion of longer term hurricane modelling, see Ranger and Niehörster (2012).

A catastrophe model for insurance works in four stages, covering (1) the hazard, in this case a hurricane; (2) the physical damage it creates, which requires modelling the vulnerability of buildings and infrastructure to wind, water, etc.; (3) insurer exposure, by looking at insurance policy terms; and (4) financial modelling of the insured losses that result. I will consider only the first component, the hazard model. This is partly for simplicity, and partly because it is the pure science part of the modelling enterprise and therefore the most likely to generalise to other examples.

There are numerous approaches to hazard modelling, so again I will describe just one. I will follow the exposition of Shome et al. (2018, p. 32), as one of the authors is based at RMS. They describe a hazard model as consisting of three interconnected modules (which can be thought of as sub-models, essentially separable components):

1. Rate module: The location and rate of "genesis events" (the beginnings of hurricanes) are modelled, in a statistical model with some physical motivations.

2. Track module: Using the genesis events from the above, the hurricane's path (speed and direction) across the Atlantic is modelled in a statistical (autoregressive) model.[6]

3. Development module: The intensity and size of each hurricane is modelled along its path, in a mixed physical and statistical model.

[6] Autoregression is a technique for forecasting time-series data which conducts a regression to predict the next time step using input from previous time steps.

Rate and track modules involve simulating hundreds of thousands of years' worth of hurricane activity, extrapolating from the historical record. The output of the hazard model that we care about is a "landfall event": an instance of a simulated hurricane moving over any piece of landmass: a Caribbean island, or part of the North American coast. The Caribbean and coastline is discretised into regions, and a landfall event is a set of variables, including windspeed and storm surge, associated with a region. So the hazard module generates a database of simulated events (calibrated to the historical data available), from which modellers can calculate rates of landfall frequency, and frequencies for the associated variables. These variables are then passed to the second stage model (the physical damage model) as inputs.

One major scientific challenge facing hurricane modellers is how to treat the process of hurricane genesis. Shome et al. (2018, p. 33) outline two broad approaches to the formation of rate models:

> *The frequency of hurricanes can be modelled as constant in time. [These] Long Term Rates (LTR) models [estimate] . . . frequencies . . . using all available data back to 1900 for landfalls or 1950 for basin hurricanes. Alternatively the frequencies can be modelled as varying in time to capture the decadal time-scale fluctuations that are observed in hurricane numbers. These. . . Medium Term Rates (MTR) models . . . are based on analysis of these observed fluctuations, and their relationships with varying sea surface temperature and climate change.*

Our focus is on MTR models. Each begins with an LTR model, and then assesses how this rate of activity will change over the next five years.[7] An important driver of hurricane formation is thought to be sea-surface temperatures (SSTs) in the "main development region" in the mid-Atlantic—and so it is SSTs which underlie the main regression relationship in MTR models. Shome et al. (2018, p. 37) display a table with the results from the 13 models in the RMS ensemble, with model names reflecting (sometimes competing) choices made in the modelling process. I will now briefly explain these names, which then allows us to describe the members of the ensemble.[8] The names, and explanations, are summarised in Table 6.

- "Direct" models use historic hurricane landfall counts as input and make a landfall prediction.
- "Indirect" models use storm formation data from the Atlantic basin to make a prediction of hurricane activity in the basin, then convert that prediction into a landfall prediction using the estimated proportion of basin storms that finally make landfall along the U.S. coastline (Jewson et al., 2007).
- "Indo-Pacific" models include the impact of sea-surface temperatures (SSTs) in the Indian and Pacific oceans on hurricane formation through their effect on wind shear in the Atlantic basin.

[7] In principle MTR models could be based on any period. RMS chose five years after taking input from their clients about which forecast period was most decision relevant (Muir-Wood and Grossi, 2008, p. 311). Five years is thought to be the right period to smooth out variability from El Niño and La Niña.

[8] As this is a proprietary model ensemble, some detective work is required here. To construct these descriptions, I compared (Hall and Jewson, 2007; InsuranceERM, 2018; Jewson et al., 2007; Sabbatelli, 2017; Sabbatelli and Waters, 2015; Shome et al., 2018).

- "Shift" models identify periods of higher or lower than average hurricane activity or SSTs in the historic data. This is due to the Atlantic Multidecadal Oscillation (AMO), and probabilities of transitions from high- to low-activity periods are estimated using historic data on tree-ring sizes, a method due to Enfield and Cid-Serrano (2006) (Jewson et al., 2007, p. 14).
- "Active Baseline" models, a mutually exclusive category with Shift, reflect an alternate hypothesis on the AMO: the low-activity period in the 1970s and 1980s was due to SST cooling induced by high atmospheric aerosol content, primarily volcanic aerosol (Booth et al., 2012). If correct, SSTs will not revert to a cool phase in the future and one should not apply a probability of shifting back to a low-activity hurricane generation phase. These "active baseline" models therefore do not include the Enfield and Cid-Serrano probabilities in their forecasts and subsequently forecast higher landfall rates than the Shift models (Sabbatelli, 2017).

The ensemble is built up by taking combinations of the above methods. It starts with 2 models: Direct and Indirect. By adding models with Indo-Pacific SSTs, we get to 4. We then add Shift and Active Baseline variants of all four—leading to 12 models. The 13th is a long-term rate model, included for comparison. RMS's long-term rate (LTR) is a statistical model based on historical landfall and basin storm data, and it models hurricane frequency as constant in time (Shome et al., 2018, p. 33). RMS's certification as a modeller for the American market (by the FCHLPM) is granted based on their LTR

Table 6: RMS Medium-term rate ensemble model names and descriptions

Rate model name	Rate model description
Long term	The long-term average calculated using historical U.S. landfall statistics from 1900 to 2013.
Direct MDR SST	Predicts hurricane landfall from the regression relationship between historical SSTs in the main development region (MDR) and historical hurricane landfall statistics.
Indirect MDR SST	Predicts basin hurricane activity from the regression relationship between the historical MDR SST and historical basin statistics. The historical proportion of basin storms making landfall is used to generate a landfall prediction.
Direct MDR+IP SST	Includes Indo-Pacific SSTs in the regression described above for the Direct model.
Indirect MDR+IP SST	Includes Indo-Pacific SSTs in the regression described above for the Indirect model.
Direct MDR+IP SST Shift	Adjusts the Direct MDR+IP prediction using the shift probability described in Direct Shift.
Indirect MDR+IP SST Shift	As for Direct MDR+IP SST Shift, but using basin hurricane data.
Direct Shift	Under the AMO hypothesis, active and inactive phases are identified in the historical data.The probability of a phase shift is calculated using the Enfield andCid-Serrano (2006) method. Combines active and inactive rates in accordance with this probability.
Indirect Shift	The Direct Shift method is applied to basin hurricane counts.
Active Baseline: Direct MDR+IP SST	Direct MDR+IP SST with the higher baseline predicted by Booth (2012) and without the AMO shift.
Active Baseline: Indirect MDR+IP SST	Indirect MDR+IP SST with the higher baseline predicted by Booth (2012) and without the AMO shift.
Active Baseline: Direct	Direct MDR+IP SST with the higher baseline predicted by Booth (2012) and without the AMO shift.
Active Baseline: Indirect	Makes a prediction of basin activity using a classification of active and in-active phases and a zero probability of shift between the two states. The basin prediction is converted to a landfall prediction using the estimated landfall proportion for the five-year periodbeing hindcast or forecast.

model and so, although RMS advertises the MTR ensemble average as providing their state-of-the-art view of hurricane risk, the LTR is often used as a reference view.

This list shows that the models included in the ensemble are not merely variants of the same model (obtained, possibly, by varying parameter values). The models fall into groups that are genuinely different, and in some cases based on incompatible structural assumptions.

4.5.1 *History of the RMS MTR*

As context for the discussion below, I want to note some history of this particular ensemble.

RMS began to offer a medium-term rate in 2006, as a response to the very active hurricane season of 2004/5 and the high losses due to Hurricane Katrina in particular. But at the start, the MTR was not the output of any model, or combination of models—it was produced by an expert elicitation process. A panel of experts was convened, and opinions elicited on whether the medium-term rate was likely to deviate from the long-term rate, in which direction and to what degree (Muir-Wood and Grossi, 2008, p. 311)

RMS moved away from the pure elicitation process as they began to develop models for medium-term forecasting. From 2007-2011, statistical models were produced and given to experts, who provided the weights for their aggregation. After 2011, the method described above was adopted.

In part this change was driven by the pursuit of objectivity: models are built on science and statistics; they are auditable. RMS describes its current technique as transparent (in the sense that the mechanism for each step can be fully described, and is presented to their clients who have confidential access to their model documentation) and based on advances in science which allowed them to move away from the less objective elicitation process.

4.6 problems with averaging in the hurricane case

Constructing a model ensemble like RMS's requires a tremendous amount of work, and the ensemble contains a lot of important information. The problem that scientists face is how to extract and communicate the information in the models to users. This is a thorny issue because it is far from clear how to interpret the (often conflicting) outputs from different models in the ensemble. Even supposing we knew this, we then face the question of how to respond to this collection of results, epistemically and practically.

A widespread and popular method is to calculate a weighted average of all model outputs, and use this average for decision-making. This works essentially as described for opinion pooling in chapter 3: models are evaluated for predictive skill using their performance on a test. There are two basic forms of test for predictive skill: forecasting and hindcasting. In a forecast test, each model makes a prediction for the same future event, you wait for the outcome, and then each model is assessed on how well its prediction fared. In a hindcast test, you take some historical data, excise a portion (the test period), and ask each model to predict what will happen in the excised time-period, using only data from before (and occasionally after) that period. In each case, a set of test predictions is scored using a particular scoring rule. The scores are then converted into weights, which set the contribution of each model: the weighted average of each model output is what gets used down the line.[9]

In this subsection, I discuss four problems with this process as it applies to the case of the hurricane models introduced above. I will also make some comments about a related, much discussed, case: global circulation models used in studies of climatic change. These problems focus on how the current decision procedure fails to provide the kind of *robustness* that one ought to want in a policymaking case. That procedure is to use this average probability as *the* probability of events in a standard expected utility decision theory. It plays the role of the agent's beliefs, and is taken as a total representation of their uncertainty.

I. Lack of basis for predictive testing

Let us start with the decision over how to test predictive performance: with a forecast test, or a hindcast test. The first problem we encounter is that neither option is suitable to an important class of cases.

Forecast testing works well for high-frequency, short-timescale predictions like weather forecasting. One can make and test predictions

[9] In the simplified example that will follow, the "down the line" usage is taking the weighted average probability as a direct input into insurance pricing decision-making. In a full catastrophe model, this probability will be fed into other modules, such as the damage module, where it is used to calculate the probability of damage above a certain value.

rapidly, and train or select for improved models. But it is not ideal for a wide range of cases. First, those in which the forecast timescale is long (where it is impractical to wait for the forecast events to occur) or the frequency low. Second, those in which the purpose of the forecast is prevention or mitigation of the event (in which case one might wait, but if the forecasted event occurs then the opportunity for prevention has been missed). Natural catastrophes have the former feature: hurricanes occur rarely, and therefore one cannot rely solely on the results of forecasts to test predictive models. Climate change has both features.

The alternative is hindcast testing.[10] This works well when we have lots of historical data for similar events, and are in a position to expect that the event-generating process is static—i.e., that the mechanism generating the events will be the same in the future as it was in the past. This justifies the basic inductive move of hindcasting, from past performance to expected future success.

There are two problems here for the hurricane case. First, the historical dataset used to score these models is small, as large hurricanes are infrequent. HURDAT2, the standard database for hurricanes hitting the Atlantic coast of the USA, is moderate in size, with about 300 storms as of mid-2018 and only 1/3 of those counting as "major hurricanes." If we split the dataset by region the numbers drop precipitously: Florida will have approximately 120 data points, Texas

[10] These are not mutually exclusive options. Typically one will update the predictive tests as more data comes in, which could be considered a form of hybrid forecast-hindcast test; especially if new results are weighted more.

65, and all other states fewer than those.[11] Shome et al. (2018) cite this paucity of data as a reason for using quasi-physical simulation models—actuaries judge that there is insufficient data to form a reliable statistical model that can be used to predict future events. Is there nevertheless enough data to build simulation models and support their predictive testing? It is difficult to say, as there are rarely definitive answers in statistics on what constitutes "enough data".[12] In order to train their models, modellers create tens of thousands of "statistical storms" to expand the dataset. This process, however, relies on the (scant) historical evidence and so it cannot remove the problem of restricted evidence.

Second, climate change may affect hurricane generation and intensity. Positive evidence for this emerged as early as 2006 (e.g., Mann and Emanuel, 2006), and some studies predict significant increases in severe storms in the Atlantic due to climate change (Bender et al., 2010). Climate change is also a common subject of post-mortem attribution studies—e.g., Risser and Wehner (2017) state that anthropogenic climate change increased the likelihood of the severe precipitation caused by Hurricane Harvey. However, the effect of climate change is the subject of fierce debate by the tropical cyclone community. A recent review reflects this uncertainty; Knutson et al. (2019) state that "opinion on the author team was divided on whether any

[11] The National Hurricane Centre has a nice summary of these statistics up to 2004 (Blake et al., 2005). As the HURDAT format is rather hard to read, interested readers may also find helpful the extensive Wikipedia List of United States hurricanes.

[12] Determining the size of dataset required for reliable inference requires knowing about the variance of the population being sampled. We do not, of course, know this about hurricane landfalls, so a small dataset might be acceptable, but assuming so is imprudent. See any good statistics reference for a discussion, e.g., Hogg and Tanis (2001, p. 386).

observed [hurricane] changes demonstrate discernible anthropogenic influence." Part of the issue is that climate change GCMs cannot resolve tropical cyclones at their base resolution (Emanuel, 2005). As resolution increases, this issue may ease (Strachan et al., 2012).

Current hurricane models for insurance do not account for climate change. As discussed above, they are built from the historical data available, with specific variations that scientists are confident in and feel able to model. The very hypothesis of climate change implies that, in future, key climate variables which drive hurricane formation will be outside of their historical ranges. What is needed is (1) evidence that these models are sensitive to the (as yet unknown) mechanisms by which climate change will affect hurricane activity, and (2) reliable projections for the climate variables that drive those mechanisms. In the absence of both, we should be wary of weighting highly a model which successfully reproduces the historical record.

It has also been argued that hindcasting is not an option for the climate case. Stainforth et al. (2007a, p. 2145) make this case, arguing that these models "cannot be meaningfully calibrated because they are simulating a never before experienced state of the system." Calibration refers to "tuning" the model—"that is, the manipulation of the independent variables to obtain a match between the observed and simulated distribution or distributions of a dependent variable or variables" (Oreskes, Shrader-frechette, and Belitz, 1994, p. 643). This is typically done by hindcasting: partitioning the available dataset for dependent variables into two parts, adjusting the model to reproduce the first part, and then testing it to ensure that it can predict the second part.

In the case of simulating climate change, the best we can do is calibrate the model against past data which, under the hypothesis of climate change, do not demonstrate the effect we are trying to simulate. As climate models are simulating a change in mechanism (a shift in climate), the events they aim to predict (various weather patterns) are expressly assumed to be different from the events in the historical record. This is a fundamental constraint of climate change modelling, and Stainforth et al. (2007a, p. 2145) argue it is so limiting that we cannot use these models to make predictions individually (or in combination in the form of an average). It also means there is no dataset on which to test different models in order to generate a skill score/performance-based weighting.

II. Choice of scoring rule

The problems discussed extensively in section 3.4, in the context of aggregating opinions, also apply here. It is difficult to discuss this in detail, however: skill scores are among the trade secrets of modelling companies and insurers, and they are therefore not in the public domain. However, from our research collaboration we do know in fact that different actors in the market use different skill scores and that these can support different results. The issues are similar to the meteorology and climate change examples discussed in section 3.4, and insurers therefore face precisely the second expert disagreement problem discussed there.

Importantly, the role of values in scoring rule selection is also present here. This undermines the claim that averaging is "objective" in the sense RMS seem to hope, and the fact that the values are buried in the technical details of scoring rule selection ought, I think, to worry decision-makers.

III. Misrepresentation of uncertainty

Writing about disagreement in the results of expert elicitation, Morgan (2014) offers another reason not to aggregate: aggregation loses information on the full distribution of responses, and focuses attention on the mean. This distracts from the extreme values, which some experts deem possible.

Morgan offers climate change as an example of where this is particularly important, quoting Oppenheimer et al.: "with the general credibility of the science of climate change established, it is now equally important that policymakers understand the more extreme possibilities that consensus may exclude or downplay" (Oppenheimer et al., 2007, p. 1505). Morgan's claim is that there are many such contexts where it is important to consider the range of expert predictions due to the nature of the potental consequences of action. I take this as an argument for reframing the problem of model disagreement. In a context where model results will be used for decision-making, we should not frame the question of model ensembles as a theoretical problem. By this I mean one for which scientific/epistemic considerations will lead to a resolution of the disagreement, resulting in an answer which *we have reason to believe*. Rather we should take the ensemble of models to be an input to our analysis of the decision-problem facing the policymaker, allowing considerations about the actions available (i.e., their consequences) to play a role in how we treat the ensemble.

Averaging often takes place before the decision-maker receives the information. They therefore don't see the spread of model results, they see only the average (accompanied, *perhaps*, by an uncertainty range for the average). This, says Morgan, is important information which is lost to the decision-maker. It tells us something about the state of our knowledge about a question. To the degree that there is spread between the model outputs, it reflects scientific uncertainty about the system and our lack of precision in modelling its relevant features. (If you were asking a Laplacean demon, who knew the true structural-dynamical nature of the world and had no measurement uncertainty, there would be no model spread. There might still be probabilities, if the phenomenon is chancy.)

Why can we not simply provide the decision-maker with both the average and some information about the spread (for example the range of outputs)? The expected utility paradigm has no role for such information: it takes a single probability distribution as input, representing the beliefs of the decision-maker. The knowledge that there are other potential probability distributions, possibly quite far from the average, is not decision-useful. What is needed is a decision theory in which there is a clear role for such information, concomitant with its importance.

Finally, weighting models relative to one another can be misleading about their absolute quality. Performance weighting merely determines how models perform on a predictive task *relative to one an-*

other. But if all models do very poorly, this procedure can generate false confidence. Stainforth et al. (2007a) argue that this is the case for global circulation models. Their claim is that the uncertainty in climate models is so severe that no one model should be considered reliable. As an example, they present the wide range of predictions for the 8-year mean precipitation over the Mediterranean basin from December–February, under a doubling of atmospheric CO_2. The range, -28% to +20%, is likely to *widen* as model uncertainty becomes better understood. Because of this empirical inadequacy, they say of any attempt to weight these models that they

> *consider this to be futile. Relative to the real world, all models have effectively zero weight. Significantly non-zero weights may be obtained by inflating observational or model variability. . . [This is misleading as it] leads us to place more trust in a model whose mean response is substantially different (e.g. 5 standard errors) from observations than one whose mean response is very substantially different (e.g. 7 standard errors) from observations. A more constructive interpretation would be that neither is realistic for this variable and there is no meaning in giving the models weights based upon it. (Stainforth et al., 2007a, p. 2155)*

In the face of such uncertainty, any single answer is highly unreliable, with not even the sign being known with confidence. It doesn't much matter which is the best of the lot, what is important is that we can't make decisions that rely on individual responses. Stainforth et al. use this to argue for decision-making only on the basis of the "envelope" of all outputs from all climate models.

This is a very specific argument, turning on the nature and extent of the model uncertainty in the case of climate models. However, it is instructive for other cases with significant uncertainty.

IV. Violation of agreement on what the right value ***isn't***

Averaging the probabilities generated by models can lead to worrying outcomes when the models have widely separated outputs, due to an agreement on some underlying causal/mechanistic factors which excludes a certain range of probabilities as plausible. Suppose that our modellers must make a choice between three mutually exclusive hypotheses, as part of the modelling procedure. Other things being equal, the choice of H_1 would make the probability low, H_2 would make the probability medium, and H_3 would make the probability high. (Perhaps they are hypotheses about the structure of the system.) Initially, let us suppose, modellers build three models: one for each hypothesis. After some empirical observations and theoretical developments, the community of scientists comes to a broad agreement that H_2 is false (or at least much farther from the truth than the others). This leaves us with two models, one for H_1 and another for H_3, generating low and high probability estimates respectively.

If our method for processing the outputs of the remaining two models is to average them, we risk generating a medium probability output. But recall that a medium probability output, on our understanding, is *only* possible under the excluded hypothesis H_2. It is tempting to regard averaging as a way of suspending judgement on the question of H_1 versus H_3. But by assumption, no such model is possible. Some decision amongst the H's is required to produce a viable model. In this circumstance a middle probability output is identical to that resulting from a choice of H_2. But all scientists agree that this hypothesis is wrong.

In this example, the experts agree on something that it is salient to preserve. But averaging isn't the kind of method that can preserve it. As we saw in the discussion of axiomatisations of pooling functions, linear pooling preserves certain agreements: identical assignments of probability to propositions. But in the case above, the probabilities are generated by underlying hypotheses, and it is agreement about *these* that we wish to preserve.

Another version of this problem arises if we consider a partition of just two hypotheses, one of which generates a low probability and another a high probability. Here averaging produces an intermediate value, for which there is no known scientific mechanism. In this case the average value corresponds to no known state of affairs; our scientific theory produces a discontinuity the probability values for this event. Note that it is not like a situation in which a continuous variable of unknown value determines the probability of an event. For example consider a coin of unknown bias. Here every value is possible, and if one does not know the bias then many argue that it is reasonable to assume the probability of heads is nevertheless 1/2. This corresponds to a particular value for the bias. In our two element partition example, there is no value for the underlying parameter which generates the intermediate value probability.

It may be tempting to think that we perform this sort of averaging all the time. Faced with two incompatible hypotheses, we assign them subjective probabilities. We then have subjective expected value for each variable of interest: the average of the values each hypothesis gives those variables, weighted by our subjective probability that each hypothesis is true. Here, then, is a way of framing my concern. Consider the probabilities generated by the models as chance-hypotheses, statements about what the objective probability is.[13] What we know is that the chance is either high or low. The question before us is whether it is legitimate, or perhaps even required, to have a moderate credence in the face of this information. If one assumes that credences must be unique probabilities then it is hard to avoid this outcome. But if one allows for imprecise credences then there is no such requirement. I will explore such credal representations in the next chapter.

Conclusion

We have a weak reason to use the average model in general, weighed against a number of problems and limitations facing averaging in the case of the hurricane ensemble. As a rough summary: averages are a reliable guide to action only when uncertainty is small (and known to be so), enough data are available for meaningful scoring, and different scoring rules produce similar results. There may be situations that satisfy these requirements, but hurricane modelling is not one of them. The nature of these problems is such that they are

[13] I don't think that the probabilities generated by these models are chances; I think they're very close to expert reports. But, as chance and expert testimony are both plausibly governed by deference principles, the analogy is useful.

unlikely to be resolved by tweaks to the aggregation methodology; a completely new approach is needed. The next chapter provides one.

5

DECIDING WITH MODEL ENSEMBLES

5.1 INTRODUCTION

I have been investigating the problem of policy-making and scientific uncertainty. In particular, I've examined the epistemic problems arising for the policymaker when experts disagree. I've shown that there are significant problems and complexities with the suite of standard tools for this problem provided by formal epistemology.

In this chapter, I shift from treating this problem as one for epistemology to regarding it as a decision problem facing the policymaker. This shift of perspective allows me to deploy new resources: the desires of the decision-maker. I do not believe that the problem of expert disagreement, as faced by a policymaker, has a satisfactory epistemic solution. The problems outlined for the epistemic approaches considered thus far arise from deep difficulties embedded in the problem. Most basic of these is the policymaker's novice status, which blocks them from being able to adjudicate technical disputes or judge expert quality in a nuanced manner. What they can do instead is to score the experts on predictive tests, hoping that these act as a suitable proxy for the epistemic goods they desire. But, as we saw in chapter 3, this process itself requires the exercise of policymaker's non-epistemic values in complex and often hidden ways.

When we approach this problem as one in epistemology or philosophy of science, we are stuck with two options that often seem unsatisfactory: trying to free ourselves from these value-judgements (and thereby often simply obscuring them) or coming to terms with the ineliminable role of non-epistemic values in an ostensibly epistemic task. The problem arises from how we have framed things. Thus far, we have conceived of the policy-making process in a linear fashion: first, the scientists present their opinions on the basis of scientific research; the policymaker (and perhaps their expert elicitation analyst) is then faced with an array of inputs where they expected one; thus, some epistemic procedure is then conducted in order to reduce this array of inputs to one; finally, the decision-making procedure continues "as usual." These scare quotes reflect an implicit assumption that this decision-making will be (roughly) orthodox: expected utility maximisation. There are two parts of this story that I now want to question. First is the strict separation between the scientific process of providing epistemic inputs to the policymaker's decision and the decision-making itself. The second is the presumption that expected utility maximisation is the decision theory in operation.

It is my belief that a sufficiently structured decision process can help policymakers decide in the face of expert disagreement, *despite* the lack of an "epistemic" resolution of that disagreement. In this chapter I outline such an alternative decision theory, which brings with it an alternative mode of engagement between decision-maker and scientist. While it is not a complete solution, I believe it provides a promising new avenue for exploring this problem.

I begin with a reconstruction of the "confidence" theory of decision-making under ambiguity, developed by Brian Hill (2013, 2016, 2019). I then turn to applying the theory to the kinds of problems I'm interested in. I show how it can be naturally adjusted to take input from the kind of model ensembles discussed in the last chapter. I then illustrate how this theory gives policymakers a tool for making decisions directly with the outputs from a model ensemble— i.e. without selecting a single probability arbitrarily or aggregating to create one. I argue that the approach does not fall prey to the problems that expert deference and opinion pooling did. As we currently lack good tools for making decisions of this sort, demonstrating the confidence approach's suitability to them is of value to policymakers and should serve as motivation for philosophers to further study the confidence approach.

5.2 DECISION THEORY PRIMER

This chapter shifts from regarding the problem of expert disagreement as an epistemic challenge and instead regards it as a decision problem for the policymaker. I begin with a brief overview of some decision theory basics, in order to provide the reader with a consistent language for the elaborations to come.

5.2.1 *Standard decision theory*

Normative decision theory considers how agents should choose when faced with decision problems. For our purposes these agents will be policymakers but the theory is much more general, simply taking them to be entities able to represent, evaluate and act upon their environments. A decision problem for an agent is a situation in which they face a choice between actions—things they are able to do, if they so choose, which have some consequences for them. The theory is normative in that it provides an answer as to what the agent ought to do, what the best approach to decision-making is.

A decision problem is a set of options (that the agent is choosing between), a set of events or "states of the world", and a set of outcomes or consequences. (All three can be modelled as propositions.) The states of the world are ways things might be, that are relevant to the choice the agent faces. The outcomes are what will happen, should she choose each of the options, if each event occurs.

Agents are represented in a decision theory as having two attitudes of interest: belief and desire. Choice is a matter of getting what you desire, given what you believe. Here is a first pass description of how decision theory represents such choices. When an agent chooses one action over another we say she prefers it, and make reference to her attitude of preference. Preference is a comparative attitude, which will be represented as a binary relation $\succsim$ on the set of actions. It is a function of beliefs about events and desires for outcomes.

The belief attitude I will take to be fundamental is *partial belief*: a comparative attitude capturing the agent's comparisons of the likelihoods of the events. This is also called degree of belief. Partial beliefs are represented primitively by a binary relation $\succeq$ on the set of states, and are also represented using a probability measure P on the same set.[1] This is precisely the subjective probability function we have seen in the previous chapters.

In a standard setting and under conditions of uncertainty, decision theory represents the agent's preferences over actions as a subjective expected utility ordering. States are assigned subjective probabilities, which represent the agent's partial beliefs. Outcomes are assigned utilities, representing the agent's preferences for what happens when they perform an act and a particular state of the world obtains. The expectation of the utility of an act is the average utility it delivers across all possible outcomes, where the weights in the average are the subjective probabilities of that outcome.

Let $\mathbb{S}$ be a partition of state propositions, and $A, B, \ldots$ be the actions the decision-maker faces. The value of an action is

$$V(A) = \sum_{i=1}^{n} u(A \wedge S_i) \cdot P(S_i|A),$$

and the agent prefers one action to another iff the first's expected utility is greater than the second's:

$$A \gtrsim B \iff \sum_{i=1}^{n} u(A \wedge S_i) \cdot P(S_i|A) \geq \sum_{j=1}^{n} u(B \wedge S_j) \cdot P(S_j|B)$$

[1] For more on the representation of $\succeq$ by P, see chapter 7.

The normative content is: the agent should choose the action they prefer, which is equivalent to that with the highest value—i.e., they should maximise subjective expected utility.

5.2.2 *Imprecise probability and decision theory*

In the standard decision theory sketched above, an agent's partial beliefs are represented using a probability measure. So, we assign to each proposition a single number between 0 and 1, representing that agent's partial belief in that proposition. Below I will advocate for a decision theoretic approach that uses "imprecise probabilities" to represent the agent's partial beliefs: sets of probability functions that generate sets of values (typically, intervals) for each proposition. This is a way of capturing features of agents' uncertainty that cannot be represented by a single probability function. Imprecise probabilities (IP) support decision rules that take such sets as inputs, as opposed to the single probability function required by the classical decision theory outlined above. This section provides a short overview of motivations for imprecise probabilities, decision rules for working with them, and some challenges thereto.

As setup to the first motivation for IP, recall that I am taking partial belief, a comparative attitude, to be primitive. The probabilities are merely representations of a more primitive mathematical object, a binary relation $\succeq$ that I call credibility. (In this chapter it does not matter very much that we strictly delineate between the attitude, partial belief, and its representation, credibility. But in chapter 7 this distinction is critical, so I introduce it here.) A "representation theorem" provides the connection between credibility and probability. This is a mathematical argument showing that the relation can be represented by a real-valued function F, where this means just that $F(a) \geq F(b) \iff a \succeq b$. Such a theorem typically specifies the form of the function, and some uniqueness conditions for it. In this setting, the relevant question is: what conditions must hold of credibility, in order for it to be uniquely probabilistically representable? The first argument for IP has the following structure: condition C is necessary for unique probabilistic representation. C is not a requirement of rationality and is sometimes irrational. Therefore, it is not rationally required that an agent's partial beliefs be uniquely probabilistically representable.

I will focus on the case where C is *completeness*. Credibility must be a complete relation if it is to be uniquely representable by a probability function: for any two propositions, either the first is more likely than the second to the agent, or vice versa, or they are equally likely. But it is no failure of rationality, for example, to lack an opinion about which of two esoteric propositions is more plausible. To use an example from Konek (2019, p. 273): I do not currently judge that it is more likely that copper's price will be greater than £2/lb in 2025 than that nickel's will be greater than £3/lb that year; nor do I judge it less likely, nor are they equally likely. Nor should I! Joyce (2010, p. 283) argues that this incompleteness is the right, even the *required* response to the scant, incomplete, imprecise, and equivocal evidence I have on the matter. It would be *irrational* to have a complete attitude in this sort of case, and in many others. If $\succeq$ is incomplete, but meets other more plausible conditions (discussed in chapter 7), it can instead be represented by a *set* of probability functions.

Second, let us consider a more straightforwardly decision-theoretic motivation. Ellsberg (1961) describes the following decision problem: An agent is shown two urns each containing a hundred balls which might be either black or red. She is told that urn A contains precisely 50 black balls and 50 red balls, while no information is offered concerning urn B. She is offered four bets, for the same amount (say, £100), that the ball drawn from a given urn will be a given colour. We can represent the four bets as AB (from urn A, a Black ball is drawn), AR, BB, and BR.

Ellsberg reports observing the following preferences over these bets: AB$\sim$AR$>$BB$\sim$BR. Under the assumption the agent is a subjective expected utility maximiser, no single probability function can rationalise these preferences. AB$\sim$AR implies that $P(AB) = P(AR) = 1/2$, as they are the only options in that case and they have the same utility. The same goes for the bets on B urn. But then it is impossible that AR$>$BR, as that would imply a higher probability for the AR than BR. Yet many find these preferences perfectly reasonable: agents have a preference for betting in an environment where they know the odds, over betting in an environment where they have no information that can be used to determine the odds. The form of uncertainty that the agent has about urn B is called ambiguity, and the preference for betting on urn A over B is called ambiguity aversion. If ambiguity aversion is rationally permissible, then it cannot be a requirement of rationality that an agent's partial beliefs be uniquely probabilistically representable.

A third motivation comes from a consideration of the distinction between the balance of evidence for an outcome, and the weight of that evidence. This latter term was coined by Peirce (1878), and popularised by Keynes:

> *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 than an accession of new evidence increases the weight of an argument. New evidence will sometimes decrease the probability of an argument, but it will always increase its 'weight.' (Keynes, 1921, p. 78)*

Put roughly, the balance of evidence determines the "best guess" probability for the event under consideration, while the weight determines how seriously we should take this estimation. As the last line of the quotation above shows, Keynes thought that the two notions operate independently.

Critics of standard Bayesian decision theory argue that a single probability function cannot capture this notion. This, they argue, is part of what goes wrong in the Ellsberg case: the balance of evidence for AR and AB is equal, *as is that* for BR and BB. But the weight of evidence is far greater in the first (Joyce, 2005).[2] Sets of probability functions, they argue, can capture weight of evidence.

A brief bit of formalism will aid with the explanation. As before, let Ω be our algebra of propositions, and let p represent a probability function $p : \Omega \rightarrow [0, 1]$—the lower case is used to remind us that now it does not represent the attitude of any agent. Imprecise probabilities, which *do* represent partial belief states, will be represented using a summary function denoted $\mathbb{P}$. The agent's attitude to X is represented by $\mathbb{P}(X) = \{p(X) : p \in \mathbb{P}\}$, where $\mathbb{P}$ is the agent's set of admissible probability functions on Ω.[3] Constraints on the probabilities of related propositions are imposed at the level of the p's, so e.g., we will preserve the logical relation between the bets BB and BR by insisting that for each probability function $p(BB) = 1 - p(BR)$. Conditioning is similarly dealt with at the level of probability functions: $\mathbb{P}(X|E) = \{p(X|E) : p \in \mathbb{P}, p(E) > 0\}$.

If we adopt this approach, then the agent faced with the Ellsberg problem has beliefs represented by such set of probability measures, which determine assignments of probability values for the bets that are admissible. For the A bets, all the probability functions in $\mathbb{P}$ agree on the probabilities of AR and AB—they are fixed by the known

[2] Precise probabilists do have responses to these challenges, which are discussed by Joyce (2005) and Howson and Urbach (2006). I will not present them here as I am merely explaining the motivation for the development of IP.

[3] Note that in the explanation above, I have been careful not to say that imprecise probability involves representing beliefs as *intervals* of probability. I make no requirement that the sets of probability functions representing beliefs are convex. This is a deliberate decision; convex creedal sets are motivated by regarding each weighted linear average of two positions as a possible resolution of the conflict between them (Levi, 1980, p. 192). However, as I have already argued, averaging is not a good solution to the type of conflicts I am interested in. The decision theoretic approach that I present is simpler if beliefs are represented by convex sets, but faces no substantial challenges if the sets are not convex.

proportion of balls in that urn. For the B bets, there is no information and so a much wider range of probabilities is permitted, perhaps as wide as $[0, 1]$.[4]

Imprecise decision-making

If we choose to represent beliefs using imprecise probabilities, we will also need a decision rule that takes sets of probabilities as input; our old rule, maximise subjective expected utility, presumes there is only one probability to take the expectation of. There are many such rules and I won't attempt a survey here. Two major approaches can be distinguished, by the decision consideration they attempt to track (Bradley, ms).

- *Caution*: In the face of ambiguity, an agent is justified in choosing cautiously, by giving greater weight to the downside risks than the upside opportunities. The classic version of this involves choosing the action that maximises the minimum expected benefit (Gilboa and Schmeidler, 1989), where the minimum is relative to the set of probabilities. Other versions introduce averages between best- and worst-case outcomes.

- *Robustness*: Agents should look for actions that achieve goals robustly, in the sense that can be expected to reach these goals under all probabilities in the set (e.g., Gärdenfors and Sahlin,

[4] As I have phrased the Ellsberg experiments, one might argue that it is clear that there is at least one ball of each colour urn B and so the probabilities are more constrained than this. We might more precisely distinguish between three cases: (1) risk, in which the proportion of black and red balls is known, (2) ambiguity, which is a case like mine above where you know the black and red exhaust the possibilities, and (3) severe uncertainty, when not even this information is given about urn B (Bradley, 2017, pp. 257–59).

1982). In some versions, a threshold of robustness is introduced that narrows the set of probabilities.

As the final piece of setup for the decision theory that I will work with, I want to note some common problems bedevilling most IP decision rules. Extreme versions of each family lead to overly cautious decision-making, and intermediate members struggle to provide non-arbitrary guidance on how to constrain that caution.

The example I gave for the "caution" family is called Maximin Expected Utility (MMEU) theory, which works as follows. Suppose that the agent has options $A, B, \ldots$, let $\gtrsim$ represent her preference relation, u be a utility function, and $\mathbb{P}$ a set of probabilities represent her partial beliefs. For each action there is an expected utility relative to each $p \in \mathbb{P}$. In this set of expected utilities, there is a minimum value representing the worst outcome that the agent thinks that action could have. One way to be a cautious decision-maker is to choose the action with the *highest* minimum expected utility.

Put formally, with $\mathbb{S}$ once again a partition of propositions representing states of the world, a "maximin expected utility" agent weakly prefers A to B if, and only if:

$$\min_{p \in \mathbb{P}} \left(\sum_i u(A \wedge S_i) \cdot p(S_i|A) \right) \geq \min_{p \in \mathbb{P}} \left(\sum_j u(B \wedge S_j) \cdot p(S_j|B) \right)$$

This is an extreme member of the caution family in the sense that it pays attention only to the worst-case scenario. This caution seems excessive if the set of admissible probabilities is the widest set consistent with the evidence available. In the Ellsberg case, or when betting on a coin of unknown bias, this will lead agents to consider very extreme probabilities and act with levels of caution that seem wildly implausible, such as preferring a bet on urn A with a very small prize to a bet on urn B with an arbitrarily large prize. This pushes us towards some other principle of permissibility for the probabilities in $\mathbb{P}$, but it is difficult to do this non-arbitrarily.

Similarly, decisions which are robust under the full range of probabilities on offer will be hard to come by—indeed, many decision scenarios will have no permissible decisions whatsoever. Instead advocates of robust decision-making attempt to define a "reasonable range" of values on the dimensions of uncertainty that make the most difference to the outcomes of the decision. Bradley (2017) reports that most approaches generate this range by assuming that a best estimate is available (typically the output of a preferred model) and generating a range around this by making small perturbations to relevant input parameters. The problem is that "sometimes the expected utility maximising option may be less robust than alternatives that are nonetheless satisfactory in terms of their expected utility. Then some trade-off between the two considerations, expected utility and robustness, must be made in order to resolve the question of what to choose" (Bradley, 2017, p. 245). A robust decision theory must answer these questions if it is to be compelling: how to generate the reasonable range, what gain in confidence is delivered by robustness over a wider interval, and how to trade off robustness against maximising expected utility.

For both families of imprecise decision rules, there is a natural and unavoidable trade-off between the confidence we gain from using a wide interval of probabilities, and the precision we desire when making decisions. We would always like to get as close as possible to the optimal decision, and our actions are typically sensitive to the actual outcome. Using a wide interval might satisfy some rational constraints, or achieve an epistemic goal like confidence, but if too little specificity is present then perhaps no action will be sanctioned (if deciding robustly) or only the most cautious (if deciding cautiously).

5.3 CONFIDENCE THEORY

At the close of the previous section I introduced the term "confidence" informally, to capture the desirable quality that decisions which are robust seem to have. The "confidence approach" to decision-making attempts to systematise this notion in a way that solves the problems just highlighted for imprecise decision rules. It it provides a systematic rather than arbitrary way of restricting the set of probability functions that are used for decision-making, and in so doing clarifies the nature of the trade-off between robustness and precision. It also builds on the "weight of evidence" motivation for IP, by providing an explicit sense in which additional evidence can increase the reliability of a probability judgement without affecting its value.

The confidence approach augments IP decision theory, by providing additional structure to determine the input to the imprecise decision rule in use. In principle the confidence approach can be used with any imprecise decision rule; for the sake of specificity in the explication below, I will use MMEU.

I will begin by presenting the notion of "confidence" and how it fits into the theory of partial belief. I will then develop the decision theory in which confidence plays a central role. My presentation of confidence is closer to Bradley (2017) than to Hill (2013), though the exposition is my own and I will note where I have added to the theory they developed. In this section, you should think of the confidence approach they way philosophers commonly think of Bayesianism: as a normative theory of rational belief and rational choice, that imputes to agents certain attitudes and requires that they fit together in a certain way to determine choices.

To get started, let us return to the Ellsberg case. The agent doesn't know the probability for drawing particular balls from urn B, and their Ellsberg preferences illustrate how this uncertainty influences their choices. Consider the probability intervals $[0, 1]$, $[0.25, 0.75]$, and $[0.45, 0.55]$. Pre-theoretically it is natural to say that the agent is fully confident that the probability of drawing a red ball from urn B lies in the first interval, is less confident that it lies in the second, narrower interval, and even less that it lies in the third. This ordinary language usage is our starting point; we want to develop this notion to serve the two purposes above, while preserving this usage.

I'll start by building out the formal machinery and then turn to its interpretation. As just expressed, confidence is a comparative attitude applied to claims that the probability of an event falls in a set. Following Jeffrey (1992), I call such propositions about probabilities *probasitions*. Formally, we model confidence as a comparative relation, denoted $\trianglerighteq$, on the probasitions. Borrowing some notation from Gaifman (1988), I use $\mathrm{pr}(X, \Delta)$ to stand for the claim that the probability of the proposition X is in the set $\Delta \subseteq [0, 1]$. So $\mathrm{pr}(X, [a, b]) \trianglerighteq \mathrm{pr}(Y, [c, d])$ means that the agent is at least as confident in the claim that the probability of X is in $[a, b]$ as they are in the claim that the probability of Y is in $[c, d]$. Each probasition can do double duty as a set of probability functions: those which make the probasition true. When the propositions and sets aren't important, I'll use π, ρ as variables for probasitions.

What properties does $\trianglerighteq$ have? Given our IP starting point, we don't want to assume that $\trianglerighteq$ is complete (that agents can compare any two claims in terms of confidence), as there may be no basis for such a comparison nor reason to make one. Reflecting on the pre-theoretical statements above, we can see that confidence must be monotonic under set-inclusion: we are at least as confident in less precise claims than we are in more precise claims. Formally, for any proposition X and two sets Δ_1, Δ_2, with $\Delta_2 \supseteq \Delta_1$, the agent must have $\mathrm{pr}(X, \Delta_2) \trianglerighteq \mathrm{pr}(X, \Delta_1)$. This is intended to be trivial: if you are quite confident that the probability of getting heads from tossing a coin is $1/2$, you must be at least as confident that it is somewhere in $[0.25, 0.75]$. You should always be entirely confident that it is in $[0, 1]$. For our Ellsberg bets, this tells us that $\mathrm{pr}(BR, [0, 1]) \trianglerighteq \mathrm{pr}(BR, [0.25, 0.75]) \trianglerighteq \mathrm{pr}(BR, [0.45, 0.55])$, just because $[0, 1] \supset [0.25, 0.75] \supset [0.45, 0.55]$.

Bradley (2017) also takes it to be $\vee$-separable, or quasi-additive: if $\pi \wedge \rho = \pi \wedge \sigma = \bot$, then $\rho \trianglerighteq \sigma \iff \rho \vee \pi \trianglerighteq \sigma \vee \pi$. Bradley is motivated in this by wanting confidence to cohere with probability in the right way: my confidence in the categorical judgement that X should be greater in my confidence in the categorical judgement that Y if, and only if, I regard X as more probable than Y.

I will not make use of confidences in categorical judgements, but the upshot of separability is that, if confidence were complete and continuous, it would determine a second-order probability function on the set of probasitions. As I am working with imprecise probabilities in part because I do not believe that rational agents need to have complete and continuous partial beliefs, it would be odd for me to take confidence to be complete and continuous. But just as precise credences are a helpful guide when theorising partial belief, second-order probabilities may be helpful to keep in mind when theorising confidence. Confidence judgements are a kind of second-order qualitative probability judgement about probasitions, themselves first-order probabilistic claims about propositions.

Let us pause for some interpretation. In the model I am describing, we have both confidences over probasitions and probabilities over propositions. What are these objects meant to represent? The agent's partial beliefs are highly incomplete: they can't compare the probabilities of the different events involved in the Ellsberg bets. (Or, if you prefer more comparative language: the agent has no judgement about which proposition is more likely.) So they don't have credences in the sense of precise subjective probabilities, and the probabilities within the probasitions don't represent credences. But nor is this simply an IP model: agents' partial beliefs are not represented by sets of probabilities, and the sets of probabilities within probasitions aren't imprecise representors. Instead, the agent's beliefs are represented by a structure called a confidence ranking: a set of probabilities *and* a confidence relation. I'll fill this interpretation out below, but at this stage it is important to note that, while my development of the approach moves dialectically from precise credences to imprecise probabilities to confidences, the objects in these theories don't have the same interpretation.

We can now introduce weight of evidence. The intuitive idea is Keynes's: confidence increases in proportion to increasing evidence supporting a claim. For a set of probability values Δ, and two distinct propositions X, Y, if the claim $\mathrm{pr}(X, \Delta)$ has more evidence supporting it than the claim $\mathrm{pr}(Y, \Delta)$ has, then $\mathrm{pr}(X, \Delta) \trianglerighteq \mathrm{pr}(Y, \Delta)$. How to weigh evidence is of course a substantive question in the philosophy of science, and I do not propose to answer it here. For the moment I assume that it will be a function of the amount of evidence (in the sense that a dataset made up of many observations contains more evidence than a dataset made up of fewer observations of the same type), perhaps alongside judgements of evidential quality. While the details are unknown, we can take it to track uncontroversial judgements of evidential support, where those exist. In particular, we can say that the judgement that "the probability of a red ball from urn A is 0.5" can be made more confidently than the judgement that "the probability of a red ball from urn B is 0.5," because of the evidence that we have about the proportion of balls in urn A.

Now suppose that we allow the Ellsberg case agent to sample urn B with replacement. We'll consider the claim that the proportion of black and red balls is equal. After 10 draws, she finds 7 black balls, and 3 red. Bayesian and frequentist statistics each have simple procedures she can follow to assess the equality hypothesis. Now suppose the number of draws increases to 100, and the proportion is now 55:45. We might naturally say that her "confidence" in the equality hypothesis has increased from when she had seen just 10 draws.

In our new language, we will say that the evidence she has gathered increases her confidence in the equality judgement. It also increased her confidence in all in those claims $\mathrm{pr}(BR, \Delta)$ for which Δ is a narrow set around $1/2$, compared to sets centred elsewhere, or those that are centred on $1/2$ but wider. Bradley puts it another way: the agent initially "has high confidence only in the unit interval. But with enough sampling she comes to have high confidence in [narrower intervals]. *So precisification of judgement occurs relative to a fixed confidence threshold*" (Bradley, 2017, p. 261).

Evidence can support wider claims, or narrower. Consider the partition of the wider interval, induced by the narrower one: $[0.25, 0.75] = [0.25, 0.45) \cup [0.45, 0.55] \cup (0.55, 0.75]$. Evidence supporting the narrower interval $[0.45, 0.55]$, such as that from the additional 90 draws, might tell us that $\mathrm{pr}(BR, [0.45, 0.55]) \trianglerighteq \mathrm{pr}(BR, [0.25, 0.45))$ while saying nothing about the relation between the outside intervals $[0.25, 0.45)$ and $(0.55, 0.75]$. Indeed, claims about $[0.25, 0.45)$ and $(0.55, 0.75]$ might be *incomparable* under $\trianglerighteq$, while logically we know $\mathrm{pr}(BR, [0.25, 0.75]) \trianglerighteq \mathrm{pr}(BR, [0.25, 0.45))$, and $\mathrm{pr}(BR, [0.25, 0.75]) \trianglerighteq \mathrm{pr}(E, (0.55, 0.75])$. On the other hand, in a different scenario those 90 further draws may have skewed the ratio of black to red balls to 75:25. Considering just our initial three intervals, this evidence supports the wider but not the narrower. Of course it will also tell us something about the subintervals, inducing greater confidence in $(0.55, 0.75]$ than $[0.25, 0.45)$.

Comparativism about confidence

In the next section I will show how we can build a theory of decision-making with confidence. But before we get there, how do we reconnect this second-order attitude with our model of partial belief?[5] Recall that I take partial belief to be a comparative attitude, best represented by $\succeq$. The incompleteness of partial belief is what generates the imprecise probability representation. And yet confidence, which is part of the belief-representation in this new approach, has thus far been discussed as a distinct attitude concerning probability judgements. More needs to be said to connect the new notion of confidence to the underlying attitude of comparative partial belief. Nothing has been said on this, to my knowledge: Hill does not have my comparativist leanings and while Bradley (I think) does, he says little to connect $\trianglerighteq$ with $\succeq$ directly.

Here is how I think of their relation. Credibility need not be complete, but it is normatively required to be monotonic over entailments, quasi-additive and transitive. We can then consider the coherent extensions of an incomplete credibility relation: all the continuous completions which preserve these rationality conditions. Each of *these* can be represented by a probability function. In IP, the resulting

[5] This section may be easier to read after reading chapter 7, which contains a detailed exposition of comparativism and my view on modelling in formal philosophy.

set of probabilities is the imprecise representor of the agent. In the confidence approach, this set gets additional structure. The agent's attitude of confidence is a comparative relation over sets of probability functions. But these probabilities are just representations of coherent extensions of the agent's credibilities. So the agent has a second-order attitude over the various ways of completing her (in-fact incomplete) partial beliefs. Instead of viewing all of these completions as equivalent—which is implicitly what happens in standard IP approaches—the agent keeps track of which completions are more or less plausible in light of their evidence, without settling on any one. These judgements are what is captured by the attitude called "confidence".

Now, each completion of credibility is a single probability measure and confidence was above defined as an attitude to sets of probability measures, or probasitions. But this is no tension: single probability measures (or more precisely singletons containing them) pick out precise probasitions. Probasitions like $\mathrm{pr}(BR, [0.45, 0.55])$ and $\mathrm{pr}(BR, [0.25, 0.75])$ represent sets of completed credibility relations, and so the judgement of greater confidence in the latter over the former can be thought of as a comparison between sets of completed credibility relations.

A potential objection to this way of thinking is that the comparativist account of partial belief is attractive because of its psychological plausibility: people can and do make comparisons of the subjective likelihood of different propositions without assigning them precise probabilities. Introducing a new attitude, confidence, which is targeted at something as abstract as a coherent extension of one's credibilities, is not very plausible. And in any case, isn't this a significant shift from how confidence was introduced above, as an attitude to claims about probability?

To diffuse this worry, we must recall that credibility is *itself* merely a mathematical representation of the actual attitude we are interested in: partial belief. What we have learned is that incomplete credibilities are not sufficient tools for representing ambiguity averse agents. Instead, we require the additional structure provided by the confidence relation.

There is no tension between thinking of confidence as an attitude towards coherent extensions of credibility, and thinking of it as a judgement about claims of probability. In truth it is a part of the agent's judgements of comparative likelihood, their partial beliefs. We model it as an attitude toward probasitions because of the role that they play in representing judgements about states of the world.[6]

5.3.1 *Deciding with confidence*

We need two further confidence-derived notions before we can develop the theory of decision-making with confidence. The first is what Bradley (2017, pp. 266-7) calls a confidence partition. The sup-

[6] There is no conflict between the confidence approach and those (rare) instances where agents do have precise subjective probabilities. Importantly, this way of thinking doesn't interfere with chance-credence principles like this one: if an agent knows the chance for some event, they should adopt it. Here is an extreme version, to illustrate the connection with the framework: if God announces that the chance of some proposition X is c, then the agent comes to have full confidence in the relevant precise probasition $\mathrm{pr}(X, \{c\})$, and no confidence in any probasitions that don't contain c. In such a case we can happily speak of c as the agent's subjective probability for X.

port of the confidence relation is the subset of probasitions on which it is complete. A confidence partition is the quotient space induced on this support by the "equal confidence" equivalence relation. The elements of the partition are equivalence classes containing probasitions that the agent is equally confident in. These equivalence classes may contain both probasitions about different events (say, if the agent judges $\text{pr}(X, \Delta_1) \equiv \text{pr}(Y, \Delta_2)$), and probasitions about the same event. The latter can occur for similar intervals, e.g., the agent might judge that $\text{pr}(BR, [0.25, 0.75]) \equiv \text{pr}(BR, [0.26, 0.75])$, as their evidence doesn't distinguish between intervals that closely. For specific decision problems we can consider sets of related probasitions about, for example, the number of red balls in an urn. These too can be used to generate confidence partitions.

Bradley then uses confidence partitions to construct a structure called a confidence ranking. This is a nested family of sets of probabilities, centred on the set in which the decision-maker has most confidence. The elements of this nesting are called "levels" of the ranking. If we start with a confidence partition $\Pi = \{\pi_1, \ldots, \pi_n\}$, with elements ordered by decreasing confidence, we can form a confidence ranking as follows. The lowest level is just the highest confidence probasition, $L_1 = \pi_1$. Higher levels are formed by taking the union with successive elements of the partition: $L_i = L_{i-1} \cup \pi_i$. Clearly this will be a nesting, $L_i \subseteq L_{i+1}$. The monotonicity of $\trianglerighteq$ ensures that $\{L_1, \ldots, L_n\}$ is a confidence-ordered nested family of sets of probability measure—a confidence ranking.

We can now do some decision theory. The core decision-theoretic insight of Hill (2013) is that the level of confidence that we require in order to act can reasonably depend on what is at stake in the decision. In addition to confidence, Hill introduces an attitude called "cautiousness" which each agent has as a feature of their psychology. Cautiousness determines how much confidence is required to make a decision, and in so doing it represents the ambiguity attitude of the decision-maker.[7] While cautiousness itself is an attitude of the agent, the demands for confidence that it generates depend on the details of the decision being made; in particular on the agent's assessment of the *importance* of the decision under consideration. Hill calls this the "stakes" of the decision.

Formally, stakes is a function from some features of the decision problem to the real numbers. The most important feature of it is that it orders the decisions the agent faces in terms of subjective importance. Cautiousness is modelled as a function from stakes to the confidence ranking; for a given stake, it determines which level from the confidence ranking should be used in decision-making. A level, recall, is an equivalence class of probasitions. As the agent regards these probasitions as confidence-equivalent, they can use the most precise probasition available. As narrower sets exclude more possibilities, agents have a pragmatic motivation to work with only the most precise set in each level.

[7] Theorem 2 in (Hill, 2013) proves that the cautiousness function is equivalent to measures of ambiguity aversion in decision theories which strictly separate beliefs and desires, such as the "smooth ambiguity" model of Klibanoff, Marinacci, and Mukerji (2005). Cautiousness can therefore be elicited using existing methods apt to such theories.

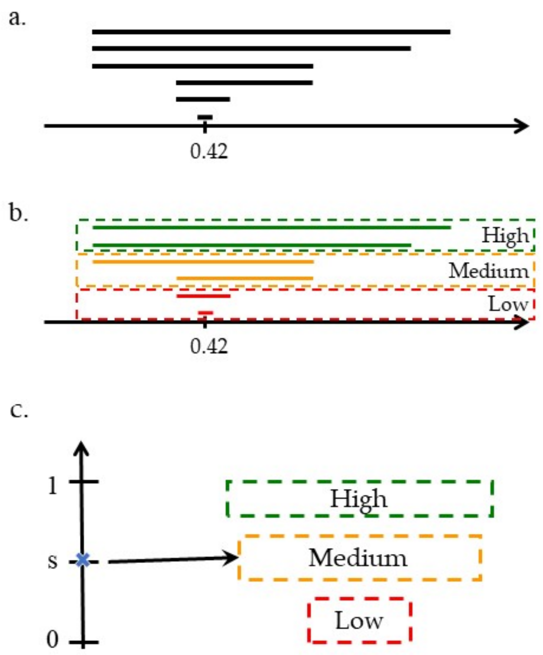
That set is then used as the input in a standard IP decision-making procedure. Which IP decision rule is used is not part of the confidence approach; it is left to the user to decide. In this way, confidence theory is meant to be neutral about the debates in IP decision theory about which rule is best. But, as we will see, it offers modellers tools for mitigating the problems associated with the two families of IP decision rules discussed above.

I will illustrate the approach with a toy example, and then highlight a crucial respect in which it needs to be augmented if it is to function well as a solution to the kind of policy problems I am interested in. I will then develop the theory more formally, and apply it to the hurricane model ensemble discussed in the last chapter.

Suppose that you are deciding whether to place a bet on your favourite drag queen, Kate Butch, winning a national drag competition. To bet you pay £50 upfront; if she wins you are paid back your £50 and receive another £50, if she doesn't win you lose your £50. So, this bet has a positive expected monetary return whenever the probability of her winning is strictly greater than 0.5. I will show how the confidence approach determines whether this is a fair or advantageous bet.

First, we represent your beliefs with a family of nested sets of probabilities. Each set represents a claim that you accept about the relevant probability, while the nesting captures the logical relationship between these claims. In our example, these claims could range from the very imprecise (indeed trivial) claim that "the probability of Kate Butch winning is between 0 and 1" to the very precise "the probability of Kate Butch winning is 0.42". Figure 2 shows such a nested family schematically.

Figure 2: a. Nested family of intervals. b. Confidence ranking. c. Cautiousness and stakes select a level.



Where do these claims come from? We can think of them as the result of the following kind of elicitation procedure, designed to work out which claims about the probability of Butch winning you would accept. We start with the widest range and narrow it in various ways to see which claims you accept. In so doing we will find the most precise claim that you will accept, which I will assume is that the probability of her win is 0.42.

I'm assuming that these judgements are based on your subjective estimate of her likelihood of winning, using the evidence you have available to you: your extensive experience in night clubs and cabaret shows, your devotion to reality TV shows about drag queens, and so forth.

If I took you through this procedure it would be natural for you to protest that you aren't very sure at all about the precise number 0.42. That would be reasonable! The confidence approach attempts to capture the attitude behind your protestations and make them part of your decision making. You will inevitably be more confident about the wider claims that centre around 0.42, and that confidence will grow as the intervals widen. In principle we could consider any number of sets, but for simplicity I will consider three: {0.42}, [0.3, 0.5], [0.2, 0.6]. In this toy example, this simple nesting is your confidence ranking. It has three levels, which we can think of as judgements you endorse with low, medium and high confidence respectively.[8] Every claim wider than 0.42 but narrower than [0.3, 0.5] is considered confidence-equivalent to 0.42 (i.e., Low confidence), and so forth. Put another way: if you saw decision-relevant differences between the intermediate intervals, you would not so coarse-grain.

How we coarse-grain is motivated by an important consideration: connecting the relative ranking of a particular decision's family of sets to a background standard of confidence. An ordinal ranking cannot say anything about "how much confidence" we have in any claim, it can only tell us how that claim is related to other claims.

[8] As a matter of logic, the full [0,1] interval is always in the highest confidence level. This could be the High level above—so that [0.2, 0.6] is judged confidence equivalent to [0,1]—or an implicit "Highest" level.

If the outcome of a bet is that I will be shot if I lose, I want to be very confident about my probability estimate; "very" reflecting the absolute importance of the decision, and not just indicating that I want more confidence in it than other estimates for the same bet. A decision-maker can do this by developing a sense of what counts as "enough evidence to warrant high confidence" and applying that standard across decision problems through the labels applied in this coarse-graining step. If there is poor evidence supporting all claims in the family, perhaps the top coarse-grained level only delivers Medium confidence. (The decision to coarse-grain to just three levels, and to call them Low, Medium, and High, is just for simplicity. One might have many more levels of evidential-support that one can discern.)

Coarse-graining to levels pegged to such a background standard of evidence allows our notion of "confidence" to decouple from the situation-specific information in front of the agent. This allows us to make decisions in a way that reflects their importance, relative to other decisions we make.

We now consider the stakes of the decision: your assessment of how important it is. How this is done can vary, but for formal simplicity, we can think of stakes as a number on a 0 to 1 scale. The term "stakes" is chosen to imply that it should be a function of the potential outcomes: as in our betting example where you stand to lose (or win) £50. There is a wide range of potential functions of these outcomes that could measure your stakes; Hill (2016) discusses their differences.[9] For simplicity let us take the stakes to be a function

[9] The 0-1 numerical form assumed here is an inessential simplification: in Hill's full presentation stakes are weak orderings of decisions. What is required for the theory

purely of the worst possible outcome—losing £50. Assessing the relative importance of a decision in which you stand to lose £50 involves reflecting on other decisions you make, the value of £50 to you, and so forth. For the moment let us assume you regard this as a moderately important decision and assign it stakes of 0.5.

Next, we must model your cautiousness. Put roughly, "more caution" means that more of the 0-1 stakes range is mapped to sets high up in your confidence ranking. In our toy example, the question to ask is "how much confidence do you need in order to make moderately important decisions, with stakes around 0.5?" Cautiousness represents your attitude to ambiguity; it is therefore subjective and will need to be elicited. Let us suppose that after such an elicitation we determine that you require medium confidence for decisions of stakes 0.5. In the case of your drag queen bet, this level of confidence is guaranteed by the interval [0.3, 0.5].

We now reconnect with imprecise decision theory, using the maximin expected utility (MMEU) rule, which mandates choosing the option that does best, in expectation, if things turn out for the worst. Note that one benefit of the confidence approach is that modellers have two "levers" of ambiguity attitude: the cautiousness function and the decision rule. Although MMEU is highly ambiguity averse, this aversion can be attenuated by the choice of cautiousness function—specifically, by choosing a function which recommends moderate levels of confidence for a wide range of stakes. (The opposite choice could boost MMEU's ambiguity aversion.) Decision-makers who are

is that the stakes relation is a weak order, depending only on the consequences of the acts (Hill, 2016, p. 85).

not completely ambiguity averse can thus still use MMEU, for instance because it is a very simple rule to implement. I use it in this toy example for precisely this reason.

You would only expect to make money betting on Kate Butch if the probability of her winning is over 0.5. As it gets lower than that, you expect to lose more and more, so because you think the probability of her winning is in the range 0.3–0.5, it is a bad bet! You don't expect to make money based on what you think the probabilities are, and there's only one way you could break even: if things turned out for the best, and the probability was at the very top of the range you think it is in. So, you don't bet.

In a single, simple decision like this, the confidence approach may seem at once overly complex and permissive. In the hurricane insurance example below, the heavy machinery will show its value.

5.3.2 *Applying the confidence approach to policy decisions*

My goal as to apply this method to the kinds of problems discussed so far in this thesis: policy decisions made on the basis of inputs from multiple models or experts.

The basic idea is straightforward. The experts/models are consulted about the state of the world that the decision-maker needs to take into account. They provide various probabilities, which are then structured into a confidence ranking. I will shortly consider a case in which an ensemble of models provides point-valued probabilistic estimates for an event. These points form the natural endpoints for interval-valued probasitions. In the drag example above, the confidence ranking is purely subjective; it reflects the structure of the agent's attitude of partial belief. In the sections to come, both the probability values and the structure of the confidence ranking will come from the scientific input. (In this way I incorporate some of the deference idea of chapter 2.) We are moving away from pure formal epistemology/decision theory and towards applied decision theory, in which pragmatic trade-offs are made to fit the framework to real scenarios, so it can assist with real decisions.

In the development of confidence rankings above I made a point of noting that the agent's confidence attitude is constant from decision to decision. Giving names like Low, Medium, and High to the levels in the confidence ranking allows us to keep track of "how confident" the agent is, in a way that goes beyond a mere ordinal ranking of the probasitions relevant to the decision in front of them. Such an ordinal ranking cannot say anything about "how much" confidence we have in any particular claim, it can only tell us how that claim is related to the small set of other claims involved in this decision; it represents *relative* confidence. But in a case where the stakes are very high (say, a threat to my life), I don't just want the best option from a set presented to me—they might all be bad. I want a claim in which I am confident *enough*, given the stakes of the decision situation.

This requirement was under-appreciated in the initial formal developments of the confidence approach by Hill (2013, 2019) and Bradley (2017). What is needed is a notion of confidence that is (at least partially) *independent* of the decision at hand. So it is important not to be misled by the toy example above: a level of confidence is not merely a coarse-graining of the nested probasitions relevant to a single decision. They represent the agent's assessments of evidential quality in an ongoing way.

If we are thinking about a policymaker's decisions, we want these assessments not only to be comparable across different decisions made by the same agent, but also to be comparable across decisions made by different agents. Policymakers typically act on behalf of a state or corporation, representing the interests of a group of people. Such decision-making is not meant to depend sensitively on the particular individual playing the policymaker role for a particular decision.

Thankfully, this demand can be met. The attitude of confidence is meant to be determined by weight of evidence: probabilistic judgements end up in a particular confidence level by being supported by evidence which exceeds a certain threshold—enough to deliver that level of confidence. The assessment of how much support a body of evidence gives to a judgement is the subject of a large scientific and philosophical literature. Confidence levels are therefore the natural point for our decision theory to connect with the philosophical and scientific discussion about evidential support. In a policy context, it will be where we make use of the literature on evidence aggregation and evidence hierarchies. These tools facilitate comparisons of evidential support, and they exist to facilitate ongoing and intersubjective comparisons in cases of multiperson decision-making.

What counts as "enough evidence to warrant high confidence" will then become a matter of policy, in the sense that rules will be established so that different people can agree on whether a particular claim can be used with high confidence. In the insurance example we will consider below, these might be set by a company policy that determines what sort of evidential backing counts as high confidence, where high confidence is what the firms cautiousness function demands for decisions involving a value at stake of more than £X million.

5.4 APPLICATION: HURRICANE INSURANCE

I will now apply the confidence approach to the case introduced in the last chapter: making an insurance pricing decision, using input from an ensemble of scientific models of hurricane formation and landfall. The example simplifies some details of actual insurance pricing by considering a very simple portfolio with only one contract, but this does not influence the philosophical points I wish to make about the treatment of outputs from a model ensemble.

An insurer wants to sell a single insurance contract on house damage due to hurricanes. They have no current contracts and plan to sell just this one, which insures against event E: "a hurricane strikes Fort Lauderdale in 2025". The contract is for a total value v =£100,000, and to simplify we will assume it is a simple binary contract, paying out either £0 if the event does not occur, or £100,000 if it does.

The insurer plans to price this contract in the tradition of Stone's (1973) constraint pricing. A company's revenue (income) can be broken into two parts: profits and costs. (If you like, these are the two places income goes: to paying costs, or to shareholders as profits.) Stone's equation demands that revenue is greater than a certain threshold, which is determined by a minimum profit level and an expectation of costs. Here is the pricing equation:

$$\Pi > \langle d \rangle + yH$$

Π denotes the premium, or price, that insurers charge—this sets their income.[10] This income, Π, must be sufficient to pay expected damages, $\langle d \rangle$, and deliver the minimum profit demanded by investors, yH. Each term will now be explained, and as expected we will see that the price for insuring some event is a function of the probability of that event.

"Damages" refers to the amount the insurer pays to its customers, which is represented by a damage function d from events to monetary values; $\langle d \rangle$ is its expectation in a given period, using the insurer's estimated probability of the events: $\langle d \rangle = \sum_i p(E_i)d(E_i)$, for some partition-forming events E_i. In our case, the events are simply $E, \neg E$ and as the contract pays nothing if the hurricane does not occur, the damages are particularly simple. Damages are a cost that any insurance company must be able to pay in order to stay in business.[11]

Insurers are required to hold capital, to ensure they can pay out claims. Regulators require this by stating, e.g., "the probability of losing more than you hold in a year must not exceed 0.5%." Hold-

[10] Ignore demand for the moment.

[11] A more realistic example would consider other expenses, such as staff costs, but I will neglect these for simplicity.

ings, denoted H, are thus a function of the "ruin threshold" k (in the example just given, 0.5%):

$$H(k) = \min\{x : p_d(> x) \leq k\}$$

$p_d(x) = p(d^{-1}(x))$ is the probability that the damage is x, and $p_d(> x)$ is the probability that the damage is above x. This is called a "loss exceedance probability" and is of critical importance in insurance pricing. The set contains all values x, such that the probability of losing more than x is below the regulator's threshold. Insurers are required to hold the smallest such value, as it is the amount required to fully cover the mandated risk.

This capital, $H(k)$, is "held"—it can't be spent or invested elsewhere—and so it is an opportunity cost to the investors in the insurance company. Investors therefore demand that the insurer generate greater profits than they expect to get elsewhere. These are calculated as the annual returns from investing the amount of capital that must be held, $H(k)$, at some benchmark rate of return, denoted y. Investors set this value by looking at capital markets and other investments they would otherwise make. It is the (opportunity) cost of capital.

This allows us to recover Stone's formula: the insurance contract's price must be sufficient to pay the expected costs of running the insurance business (damages) and deliver the benchmark profits expected by investors.

Our simple example involves selling just one contract, which turns out to be a bad way to go into the insurance business. The formula for $H(k)$ means that so long as $p(E) > k$, the insurer will need to hold the full contract value ($H = v$). So if it turns out that, say, $p(E) = 0.01$, the ruin threshold is $k = 0.005$[12] and the cost of capital is 5% ($y = 0.05$), Stone's equation says that the minimum premium is £6,000. The ruin threshold is set by a regulator, and the value of y is dictated by capital markets, so the insurer's problem is to determine $p(E)$, given the results of the model ensemble.

Our scientific input will come from a toy ensemble, that captures the salient features of the RMS ensemble from chapter 4. Our scientific modellers construct ten models, $m_1, \ldots, m_{10}$, which encode different views about, e.g., how hurricanes move across the Atlantic, and how the factors influencing their generation will turn out in 2025. As the details will not matter here, we will not describe how these models work except that they generate $p(E)$, and that one of them—m_8—was built for a different region but works for Florida. Table 7 shows ten numbers that we will use as our model outputs. The "standard" approach would be to score these models on their predictive skill, as described in chapter 4. Suppose that we have done this, using a popular scoring rule R. The normalised scores and outputs for $p(E)$ are shown in Table 7.

Let us consider how the standard, averaging approach would price the contract. Using superscript A for weighted average, $p^A(E) = 0.0072$ and the expected damage is $\langle d \rangle^A = p^A(E)d(E) = 0.0072 \times 100,000$ =£720. The required holdings are H =£100,000. If we take

[12] This is a realistic ruin threshold for major reinsurers like Swiss Re, equivalent to expecting to go bust once every 200 years.

Table 7: Toy ensemble model outputs

Models	$p(E)$	Weight (%)
m_1	0.007	23.7
m_2	0.0071	20.7
m_3	0.0068	15.8
m_4	0.0074	11.6
m_5	0.0076	11.5
m_6	0.0061	7.3
m_7	0.0083	3.2
m_8	0.0086	3.0
m_9	0.0091	1.7
m_{10}	0.0092	1.6

the cost of capital to be $y = 0.05$, then we have the following minimum price

$$\Pi^A > \langle d \rangle^A + yH = £5,720.$$

This assumes that we proceed by straightforwardly using the average model probability. But in chapter 4, I argued that we should not do this, by providing a number of reasons to doubt the epistemic and practical reliability of the ensemble average. Some of these worries are well-known to insurers, in particular the arbitrariness of the choice of scoring rule and the misrepresentation of uncertainty. Indeed the insurers that we worked with on this case study were motivated to approach us precisely because they feared that their current method—averaging—was not duly accounting for uncertainty. What insurers lack is a systematic way of factoring in the uncertainty that they think is missing from the ensemble average. Concerned that they will under-price their contracts and expose themselves to ruinous risk, skeptical underwriters today apply their judgement based on experience insuring hurricane risks to “scale up the risks.” In practice this means multiplying the average event probabilities by some $\alpha > 1$. They have, however, no principled way of determining the value of α, and its value is set by people whose expertise is in insurance underwriting rather than scientific modelling.

This introduces a second target for my discussion here. To represent it in our toy example, I will assume that the insurer doubles the aggregate probability $p^A(E)$. Going through the calculations with that probability, we get the “safety” price $\Pi^S > \$6,440$, which is much higher than the “technical” price Π^A. My goal is to demonstrate that the confidence approach can avoid the issues raised for averaging, and provide a better way of incorporating additional uncertainty than “safety pricing.”

To apply the confidence approach we must formulate a set of probasitions, claims about $p(E)$. Here, I will use a very a simple method of doing so for illustration; I will discuss more elaborate alternatives in the next section. We will assume that scoring rule R has reliably identified the best model, m_1, and build intervals around it. Our “lowest”, most specific claim is that $\mathrm{pr}(E, 0.007)$. We form wider intervals by including predictions in order:

$$\begin{aligned}
\Delta_1 &= 0.007,\\
\Delta_2 &= [0.007, 0.0071],\\
\Delta_3 &= [0.0068, 0.071],\\
&\ldots\\
\Delta_{10} &= [0.0061, 0.0092].
\end{aligned}$$

Due to their nesting, we know that $\mathrm{pr}(E, \Delta_i) \trianglerighteq \mathrm{pr}(E, \Delta_{i-1})$. We now examine the weight of evidence underlying each probasition to enrich these confidence judgements. The goal is to coarse-grain by deciding on which probasitions to group as confidence-equivalent, thereby forming a smaller number of levels. As our insurer has made no decisions of this sort before, they do not have an ongoing assessment of evidential quality and so construct the confidence levels using only the model outputs. For simplicity we can suppose that these ten models represent all the major relevant scientific views. As a first pass, the insurer could decide to coarse-grain by model support: using up to 4 models will yield Low confidence, 5–7 Medium, and 8–10 High.

However, in consultation with the modellers they note that this would result in the narrowest set in the High level being Δ_8. But the modellers doubt m_8 as it was built for a different region. In the preliminary confidence ranking above, a shift from Medium to High confidence is brought about by including m_8, but the insurer now thinks it doesn't carry enough evidential weight to justify a shift in confidence level. So the insurer revises the coarse-graining so that Low involves models 1–4, Medium 5–8, and High 9–10. In essence, they decide that adding m_8 has no impact on confidence, $\mathrm{pr}(X, \Delta_7 \cup m_8) \equiv \mathrm{pr}(X, \Delta_7)$, while adding m_9 brings about a significant change in confidence, $\mathrm{pr}(X, \Delta_9) \triangleright \mathrm{pr}(X, \Delta_7 \cup m_8)$. The three levels in the confidence ranking are shown below, and illustrated in Figure 3.

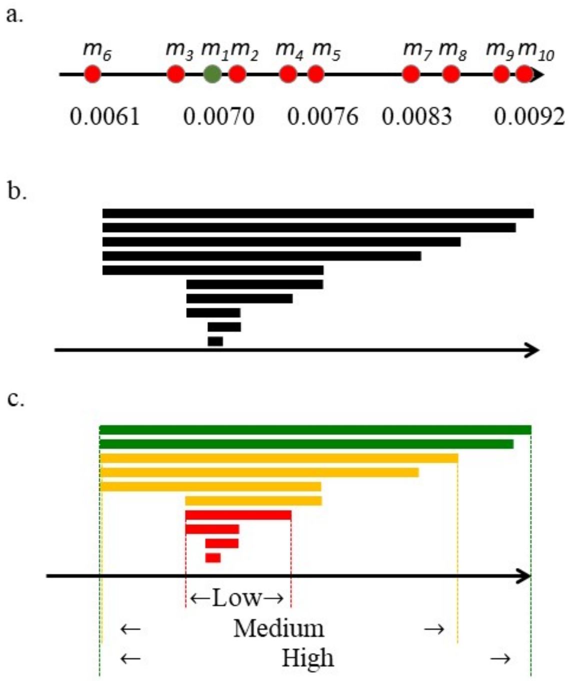
$$
\begin{aligned}
L &= \{\{0.007\}, [0.007, 0.0071], [0.0068, 0.0071], [0.0068, 0.0074]\} \\
M &= \{[0.0068, 0.0076], [0.0061, 0.0076], [0.0061, 0.0083], [0.0061, 0.0086]\} \\
H &= \{[0.0061, 0.0091], [0.0061, 0.0092]\}
\end{aligned}
$$

This is a simplified example of how scientific facts about the models inform confidence level formation. In real cases, the process of forming confidence levels may be quite complex. For example, the decision-maker and scientists might look at the evidence used in the construction of each model: different evidential bases (perhaps due to the scientific disagreements generating the ensemble) can generate different incremental gains in confidence. Suppose, for example, that one model makes use of an extensive dataset that another does not and cannot. Ceteris paribus, this is a reason to weight the result from the first model more than that from the second.

The tool we have for representing this in our model is basic: we can judge a probasition to warrant strictly greater confidence than another, or judge two probasitions to warrant equal confidence. But this does allow us to form the confidence ranking in a manner that reflects scientific judgements of evidential weight (which, recall, I take to include evidential quality).

Note an important feature of this approach: the confidence levels reflect the *policymaker's* coarse-graining of the probasitions. They correspond to the decision-maker's ongoing assessments of what counts as evidence sufficient to warrant high confidence, for example. Now the policymaker obviously cannot conduct this coarse-graining independently, for they don't understand the evidence that underlies the probasitions. So the formation of the hierarchy of nested sets must be a collaborative process. The policymaker and scientists must work together to map the probasitions relevant for *this* decision into a confidence ranking that reflects a background standard of evidence that is suitable to all the decisions the policymaker faces.

Figure 3: Toy output from the hurricane model ensemble. a. Probabilistic outputs from models. b. Nested sets $\Delta_1, \ldots \Delta_{10}$. c. Confidence ranking.



How can we characterise the stakes facing this insurer? This contract will constitute its whole business and so the risk of ruin is high.

Still, no one's life is at stake and there is no impact on anything outside of the realm of this decision (no other business which might be taken down). So, the insurer concludes that their stakes are moderately high, which we will represent with $s = 0.75$.

Next, we describe the insurer's ambiguity attitude (which corresponds to cautiousness). As insurance of natural catastrophes involves significant ambiguity, it seems reasonable to assume that this insurer is not overly ambiguity averse. Here is one cautiousness function which exhibits only moderate ambiguity aversion:

$$\mathcal{D}(s) = \begin{cases} L, & s < 0.6 \\ M, & 0.6 \leq s \leq 0.9 \\ H, & s > 0.9 \end{cases}$$

Applying this to the example outlined above we see that the decision-maker resolves to use level M. Recall that the insurer regards the sets within each level as providing equivalent confidence, and so will make decisions using the narrowest interval available at a level, viz. $\Delta_5 = [0.0068, 0.0076]$.

We can now apply our chosen decision rule. In insurance, higher probabilities represent worse payoffs for the insurer, and so MMEU uses the highest probability in the range: $p^C(E) = 0.0076$, with superscript C for confidence. We therefore have the following expected damages $\langle d \rangle^C$ =0.0076×100,000 = £760. The holdings are exactly as before. We therefore get $\Pi^C > \langle d \rangle^C + yH =$ £5,760.

Comparing the lower-bound value for Π^C, £5,760, with the crude "safety" price Π^S, £6,440, we see that the confidence price is 10.5% lower. This is a very large difference when pricing insurance portfolios. It will price out many customers, and therefore represents income foregone due to arbitrarily determined caution.

Comparing Π^A and Π^C, we see that the confidence approach recommends pricing the same contract (at least) 0.7% higher than the averaging approach. This small difference is an artefact of our toy example: selling only one contract imposes exceedingly high capital costs, as we required the insurer to hold the entire contract value as capital. If we adjust the example to have the insurer sell 20 contracts, holding £100,000 to cover all of them, and spread the capital cost evenly among them so that $h = H/20$, the prices would be as below.

$$\Pi^A_{20} > \langle d \rangle^A + yh = £970$$
$$\Pi^S_{20} > \langle d \rangle^S + yh = £1,690$$
$$\Pi^C_{20} > \langle d \rangle^C + yh = £1,010$$

Now, the confidence price is 4.1% higher than the aggregate price, and 40% below the "safety" price. Our structured approach to uncertainty classifies a number of sales as imprudent which would go ahead under the average price, but far fewer than are excluded by the ad hoc safety price. The lesson from this toy example is that the current "rule of thumb" uncertainty management is not only baseless; it is not cost-effective.

Note also that the stakes and cautiousness functions, while subjective, are stable attitudes of the decision-maker that persist across decisions. The confidence approach therefore ensures consistency across sets of decisions in a manner that ad hoc uncertainty management cannot.

5.4.1 *Conclusion*

The standard approach to working with model ensembles is beset with problems. Aggregation relies on a non-unique predictive test and scoring rule, whose choice is difficult to motivate to decision-makers. It requires significant data, which may not be available. Crucially, it misrepresents the state of scientific knowledge to decision-makers by producing a single value for $p(E)$, without reflecting the underlying uncertainty. This is compounded by decision-makers not knowing what to do with uncertainty information, were it to be given to them.

In the confidence approach we are as explicit as possible about uncertainty at every stage. Decision-makers are presented with a variety of options: different sets of probabilities, each with an attached "cost" to their use in the form of the confidence it can support. One can always demand more specificity, but it is clear what is given up when doing so. There is a natural, and I think valuable, link between the importance of the decision, the confidence that importance demands, and the formulation of decision-input.

In our insurance case-study the benefits are marked. Insurance is meant to put a "price on risk" so that people can pay protect themselves from the unexpected. Insurers would also like to put a price on the kind of uncertainty that we discuss here: not knowing what the probability of some event is. The current options available to insurers are all ad hoc and there is no guarantee for insurance companies that their staff are responding to different risks (hurricane, earthquake, wildfire) in a common and systematic way. There is also no guarantee that identical risks (i.e., events for which the ensemble outputs are identical) will result in identical prices, if the underwriter judgement is applied contract-by-contract and is not constant. (Recall that our "double the probability" representation of ad hoc adjustments is a simplification; what insurers reported to us is simply that underwriters adjust the price proposed for a contract using their judgement. The important part is the fact that it is done at the contract level.)

This approach allows insurers to systematically set this kind of "uncertainty premium." For an insurance company, three of the ingredients discussed above will be a matter of policy. They will need to agree on a way of measuring stakes that allows them to compare the different kinds of decisions they make and decide which are more and less important. Cautiousness can similarly be determined by a high-level decision about how much confidence to demand for decisions of various stakes. Finally, a decision rule will need to be selected; either maximin expected utility or one of its competitors.

With these three elements in place the cautious confidence approach provides a recipe for pricing insurance that is sensitive to all of the evidence available for that decision and which responds naturally (through the cautiousness and stakes) to the different nature of each decision taken. This kind of flexible but systematic treatment of uncertainty is what insurers tell us they have been missing in catastrophe insurance. The averaging procedure it replaces has the superficial appearance of objectivity because it is conducted by scientific specialists and does not appear to involve non-epistemic values in the manner that the confidence approach does. But, as we saw in chapter 3, there are value commitments involved in the choice of scoring rule. Better, I think, to make them explicit and provide a structure that helps decision-makers see how they interact with other ingredients in the decision-making process.

The last two paragraphs also demonstrate a second benefit: the confidence approach fits naturally with the kind of distributed decision-making and corporate responsibility found in large insurance companies. The different parts of the recipe are naturally provided by different stakeholders. The cautiousness function is ultimately determined by the shareholders' appetite for uncertainty. The way of determining stakes will be set by senior management in charge of portfolio management and capital allocation. The probability functions themselves, along with their nesting and grouping into confidence levels, come from the science department who cover that particular risk. This is a better fit with how insurance decision-making should work than having underwriters "adjust" scientific estimates of probability individually.

The approach outlined here is not restricted to insurance or hurricane modelling. In principle, the approach can be expanded to cover any decision-support using a model ensemble—including non-probabilistic outputs. Doing so would better reflect uncertainty and strike a balance between cautious decision-making in the face of uncertainty and avoiding complete decision-paralysis.

Before continuing, let us pause here to consider a potential objection: While the ad hoc safety price is obviously unjustified, the defender of averaging might protest that this shouldn't be the metric of success. In particular, in cases where the average has greater "skill" than the best model, the confidence approach appears to result in the use of a probability-value that is generated by a less skillful model. Surely the decision-maker should simply rely on the average? In reply I say: decision-making using the average is overly reliant on the scoring process—which is clear from use of the term "skill" in my presentation of this example. As discussed in both chapter 3 and 4, there is a myriad of scoring rules and a degree of arbitrariness in the choice of which to use. In addition, in this particular case we suffer from a paucity of data, that undermines the reliability of any measure of "skill." The confidence approach mitigates these worries by introducing a flexible degree of robustness. As I will discuss below, there are different ways one might construct the nested sets and it may well be reasonable to centre them on some sort of average. But the key to the confidence approach is that, as the stakes increase, one uses wider intervals around the centre, thereby guarding against concerns about how that centre was identified. So even when the average has greater "skill" than even the best individual model, one should prefer the confidence approach to using simply the average.

5.5 METHODS FOR CONSTRUCTING NESTED SETS

The simple implementation of the confidence approach to model outputs described above is by no means the only option.[13] Our view is that there is no "one size fits all" method for the construction of nested sets, given the diversity of target systems and modelling endeavours. Instead the set-construction method will depend on the specifics of the ensemble. In this section we make a start on a "toolbox" for model-based decisions: outlining several potential set-construction methods and identifying what each requires of the ensemble and when they are likely to perform well.

In the toy example we constructed our intervals by starting with the best model, m_1, and including the next best (according to rule R) each time. But we could also describe what we did as starting with m_1 and including the next closest model each time, with respect to the Euclidean distance between outputs. In the toy example these procedures generate the same result, and we did not specify which we were following at the time. But in general, we may not have a reliable rule, and these two orderings may diverge. We now outline a decision-tree for how to construct a nested hierarchy in the general case.

The first question is: can you identify a model output or outputs to act as a centre for the nesting? (We will consider different ways you might successfully accomplish this in section 5.5.1) If yes, then: do you also have a reliable ordering of model outputs? If yes, then we

[13] This section includes material from a co-authored paper with Roman Frigg and Richard Bradley. I contributed ≈60% of the work in this section.

recommend forming the nesting in line with this ordering. But recall that, in chapter 4, I outlined various problems with scoring models in the hurricane case—at this point in the decision-tree, our case likely yields a "no". In that case, I recommend forming the nesting by including models in distance order.

If you cannot identify a centre (5.5.2), then we ask: can you defend the use of one of a suite of statistical methods which construct a centre (and a nesting to go with it)? I will consider one example: using an equal-weighted mean as the centre, and the central intervals of a Gaussian distribution to define confidence levels. If no such option is suitable, then you are in the worst-case scenario and must use only the widest envelope of your model outputs.

5.5.1 *Cases with an Identified Centre*

Let us consider cases where we can identify one model as best, and so we use it as the centre. First, this identification might use a skill score, or multiple scores. Recall that in chapter 4 I presented a number of limitations of using the weighted average of the hurricane ensemble outputs, two of which also speak against the use of a scoring rule to rank models: (1) there are many such rules and choosing between them is a complex matter over which experts disagree, and (2) there may be limited data for testing, in which case the scores may be unreliable.

In the simplest case when neither of these problems is salient, we will have a single scoring rule which makes use of sufficient data to identify one model as best. If so, we use it to form the centre. Note that we needn't always centre on a point output. In situations where we are uncertain, it may be natural to have the most precise claim we are willing to accept be interval-valued: an uncertainty range around the best output, reflecting the uncertainty in even our best model. More complex cases will involve multiple plausible scoring rules. If they agree on the best model, we not only have a starting point for the hierarchy but can regard it as having a degree of robustness. If the rules disagree, we are in a difficult situation in which there are multiple best models (Betz, 2009). In such a situation we can still follow the robustness thought and form a central interval from the best model identified by each scoring rule. Finally, we may also have a method of identifying a centre that doesn't rely on a skill score, for example if experts tell us one model is best without providing a performance-based rationale. (The same considerations discussed for scoring rules apply.)

Given that we have a centre, we now need to form the nesting. Here the natural question is: can we form a reliable partial ordering of models, reflecting their strength? We consider first the positive answer case, then the negative.

5.5.1.1 *Nesting Using a Partial Ordering*

As one of the main ways of identifying a centre is using a skill score, we will first consider the case where we trust that rule (or rules) to partially order the models. As with the centre, good cases using a scoring rule (SR) order are those where there is a natural rule and plenty of data. Here the rule's ordering gives evidence of model strength, and we can follow it as in the toy example. If there is more than one scoring rule on the table, we can attempt to form an SR order by consulting each of them. In the best case, they agree and we use the resulting order. This would confer some robustness upon the ordering, and the resulting hierarchy. If they disagree, we are back in a difficult case. Following the thinking above, we might try to form the interval about the centre by including all the second-best model outputs, and so on. This is a rather cautious approach, and relatively small differences between the rules could lead to a very coarse-grained hierarchy.

A less cautious approach is to break the tie between, e.g., two models each ranked second by some scoring rule, using the distance of each output from the last interval in the nesting. This produces a finer-grained hierarchy, which may be helpful when the SR order is too coarse to allow for the desired number of confidence levels.

5.5.1.2 *Nesting using Distance*

If we have no reliable ordering information about models, other than the identified centre, then we can use the distance of models from the centre to form a naïve ordering. A hierarchy built on this ordering will respect the logic of confidence and will produce relatively fine-grained hierarchies (unless many model outputs happen to be equally spaced), which can then be coarse-grained to form confidence levels. This method is conservative, in that it uses only model outputs to form the hierarchy, unlike methods discussed below.

The problem with it is that distance-ordering needn't track any facts about model strength. When we use an SR order, we know something about the confidence gains resulting from moving to a wider interval: each step up in the hierarchy delivers weakly less incremental confidence than the previous step. Using a distance-ordering does not ensure this, and so the resulting hierarchy is less informative. This makes sense in our more uncertain case, but it is why we do not endorse distance ordering when there is a defensible SR order available.

5.5.2 *Cases without an Identified Centre*

We now consider cases where we can't identify any centre. Here the only facts available to a decision maker are the model outputs themselves; we are in a case of more severe uncertainty and can use only distributional properties of the ensemble to generate our hierarchy.

5.5.2.1 *Nesting using Statistical Methods*

Although we cannot identify any model in the ensemble as best, we may nevertheless be able to construct a centre for our hierarchy. The thought here is that the ensemble contains useful information about the phenomenon of interest, at the level of individual model outputs, but that we are unable to extract it through model comparisons like performance testing. Treating the models statistically, we can attempt to structure this information at the level of the ensemble and use it to guide our decision-making.

There are many statistical methods and comprehensive discussion of their uses in the context of the confidence approach is a project for future research. We will here briefly outline a simple method that utilises point estimates, from a natural science setting closer to our case study: the Coupled Model Intercomparison Project (CMIP5) for Global Circulation Models (GCMs). The foundation of this approach is "one model, one vote" (each model is treated equally), with results generated by simple statistical analysis. To begin, we calculate a straightforward arithmetic mean of model outputs, $\bar{m}$, and use this as the centre of the nesting (Betz, 2009, p. 754). We then calculate the variance of the output set, defined simply as $s^2 = \sum_i (m_i - \bar{m})^2/n$. Assuming error is Gaussian, one can then input these into a Gaussian distribution $G(x) = c\exp[(x - \bar{m})^2/2s^2]$, where c is a normalisation constant. With this in place we can calculate nested intervals directly from the distribution. We can centre on the mean, and then consider various centred intervals of the distribution: the central 50%, central 80%, etc. These form the sets of the nested hierarchy.

This method has limitations. The key assumption here is that all models are of equal value—this underlies the simple arithmetic mean and uniform variance analysis. This may seem implausible, either because not all models are on a par, or because they are not independent and so "voting" may double count—see (Knutti, 2010). The centre is also sensitive to the number of models in a way that scoring approaches are not: the addition of duplicate models may move the centre without adding additional scientific information. This approach is therefore best used in situations where there is a fixed and small number of models, no method to rank them, and all of their output values are plausible—a description many climate scientists believe holds for GCMs.

Statistical methods are also common in economics (where the term "model" is often used differently, to refer to a function of the underlying data), for instance in the robustness method of Hansen and Sargent (1982). We will not discuss the large range of options available in this case—including maximum entropy, Bayesian model averaging, and so on. These methods typically utilise richer information than we have presented in this paper—such as a full probability distribution rather than merely a point estimate. The confidence approach works with each of them, and at a high level of abstraction the procedure is the same: centre the hierarchy on the constructed central estimate of the relevant probability, and then form confidence levels using distributional facts.

5.5.2.2 *Working without Nesting*

In the worst cases, we will not be able to rely upon any of the foregoing methods. We may not believe any scoring rule can adequately measure model skill, be unable to identify a best model or models, and have reasons to doubt the applicability of distribution-fitting or other statistical techniques.

Stainforth et al. (2007a) argue that this is the case for GCMs in the CMIP5 ensemble. They argue that today's GCM ensembles provide only a "non-discountable envelope" of outcomes—i.e., a set of possible outcomes. No individual model can provide a reliable central estimate, and therefore the ensemble should not be used to create one through aggregation. Any construction of a PDF, such as through the method described above, is therefore likely to mislead decision-makers through false precision (Stainforth et al., 2007a, p. 2158). Worse, they provide only a lower bound on the range of uncertainty, because further uncertainty exploration is likely to increase it (Stainforth et al., 2007b, p. 2166). This is an extreme view—if it were widely accepted, the IPCC process would not be seen as generating anything of decision-relevance—but it is a useful limit case when considering the options within our approach.

Stainforth et al.'s arguments for these conclusions are complex, but at heart the issue is multiple uncertainties, each severe and in combination so limiting that we cannot use these models to make point predictions. The members of the ensemble are so interdependent, they argue, that we should also not believe that model agreement lends any additional confidence. All we can present is the range of results generated by our models, and the range of uncertainties accompanying them. These are useful: they represent informed assessments of possibility, formulated by our best experts. They therefore determine a region of output-space that is "non-discountable"—i.e., that we should not expect the truth to lie outside.

In situations like this, where the ensemble is thought to represent only a part of our uncertainty and where the model results are not particularly reliable, what can the confidence approach say? We could follow the recipe of one of the statistical methods above to form a hierarchy, and therefore provide some sense of more- and less-confidence-generating claims. But, when we coarse-grain to confidence levels, even the widest set in the hierarchy must be regarded as having Low confidence—where this is now interpreted in the sense of being "non-discountable." In order to gain more confidence, we must move to yet wider sets and here we may have little to guide us. The confidence approach tells us that if our decision is high stakes, and our cautiousness dictates High confidence, we will have to use some wider interval than any supported by the model ensemble (in the extreme, including [0,1]).

An additional problem facing decision-makers is that there may be serious possibilities that are not reflected in the range of model outputs, and in such a situation it is unclear why the envelope of the model results can be seen as narrowing down the non-discountable option (Betz, 2013). The IPCC recognises this possibility and in response has endorsed the practice of "downgrading" prediction confidence. Here, results that are generated by examining the 5-95% range of model results (for instance, for global mean temperature change in 2100, under a particular forcing scenario) are reported as merely "likely" (>66% probability) rather than "very likely" (> 90% probability) (IPCC, 2013, Table SPM.2). This way of catering for the possibility that something that the models do not simulate happens uses expert judgement (Frigg, Thompson, and Werndl, 2015, p. 973). Insofar as this reassignment reflects information that scientists hold about limitations in prevailing modelling, it is surely more transparent to reflect it through the confidence grading of different probability ranges than by downgrading the probabilities themselves, e.g. by reporting the results are "very likely" at medium confidence, but "likely" at high confidence (see Helgeson, Bradley, and Hill, 2018; Mach and Field, 2017).

5.6 objections and concerns

This completes my presentation of the confidence approach and how can be applied to decision support using the results of the model ensemble.[14] Like any approach, it has pros and cons. In this section I want to turn to some objections and concerns that have been raised for the confidence approach. As the literature on this approach is not very large, these are concerns that have been raised to me in conversation when I presented this work. Below, I introduce objections in italics and then answer them or indicate where further work is needed.

This method selects a boundary-point of one of the sets in the confidence ranking. Doesn't this mean we are effectively making a decision with one of the models that we know to be less good than the model which is at the centre of the ranking? This is a misunderstanding that has arisen in presentations of my hurricane insurance example. It arises because of the simplicity of MMEU, and the nature of the insurance decision studied. In the way presented the example it does appear that I select a probability: specifically, in section 5.4, I said: "in insurance, higher probabilities represent worse payoffs for the insurer, and so MMEU selects the highest probability in the range: $p^C(E) = 0.0076$", which is indeed the output of one of the models (m_5). But what has happened

[14] From this point onwards, all material is my own work once more.

here is not like selecting a single model's output on the basis of a scoring procedure, for either belief or to make decisions by subjective expected utility maximisation. MMEU is a method for selecting an *action* that performs best according to its criterion. In this case that action is a price, which is a function of the event probability. In this sense, the example's strengths (the models output probabilities, the decision is clearly probability dependent) may also be weaknesses (the probability is required to specify what the action is, leading to the misleading impression that we are primarily interested in determining which probability is "best").

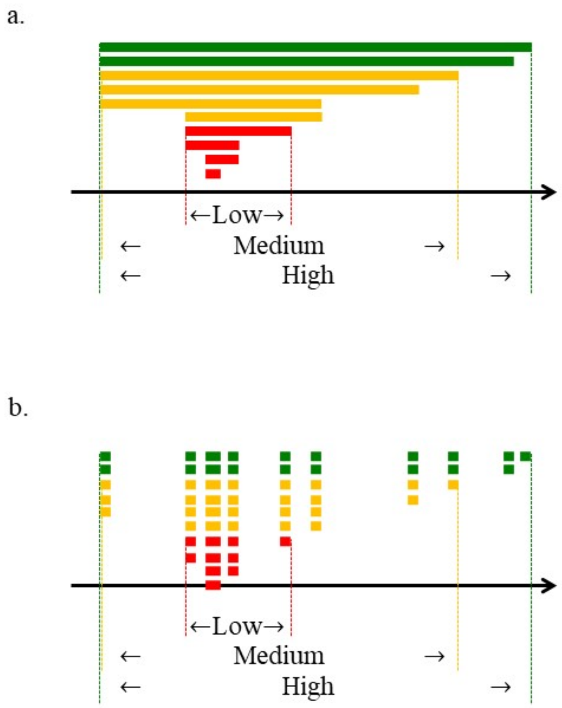
Furthermore, there are conditions under which we will make the decision using the model that is at the centre of the ranking. That is precisely when our decision requires the level of confidence that can be guaranteed with that model alone—in our example, Low confidence. But when the stakes are higher, more confidence is required and we must look to wider sets to secure it. So our situation isn't one where we are relying on a less good model when we could rely on the best model. We are in a situation where we *can't* rely on just the best model, and so we have considered a wider set.

You argued in section 4.6 that we should not average model outputs from this ensemble, as we may end up with a value that no model endorses and which all deem impossible. Yet, for certain decision rules your confidence approach will also select such values. For instance, if you select the set $\Delta_5 = [0.0068, 0.0076]$ *and use a rule that uses an arbitrary mixture of the best and worst case,* $\alpha 0.0076 + (1-\alpha)0.0068$*, then for many values of* $\alpha \in [0,1]$ *you will use precisely such a value.* It is true that, as I presented this example many ambiguity decision rules will use arbitrary combinations of model probabilities. But this is a feature of a non-essential simplification that I made in presenting the example, rather than a feature of the confidence approach itself. I chose to have the sets Δ_i be convex, i.e. intervals in the real line. But nothing in the approach requires that we do this. The sets in the nesting can contain only the point valued outputs of the models, or balls around those outputs (representing initial condition and parameter uncertainty), without including all of the points between those outputs. Figure 4 shows a confidence ranking made with such "gappy" sets.

There are too many "levers" for ambiguity attitude in this approach. There are some in the decision rule (i.e., MMEU is ambiguity averse), some in the cautiousness function (which is intended to represent that attitude) and perhaps even some in the stakes function (which in the example considered only the worst-case scenario)! It is true that the decision rule and cautiousness function jointly represent ambiguity attitude. This was highlighted as a benefit of the approach earlier on, where I noted that it allows us to use particularly simple decision rules (MMEU) and then to "dial down" their ambiguity aversion. This flexibility does call for a principled approach to determining how and where to represent an agent's ambiguity attitude.

Suppose that we can measure ambiguity attitude as a single parameter, $\alpha \in [0.1]$, as in the Hurwicz (1951) criterion. This parameter is typically thought of as providing a way of choosing a point on the spectrum between maximum pessimism (MMEU) and maximum optimism (maximax expected utility). But in our approach, we can se-

Figure 4: Confidence rankings for the hurricane ensemble. a. Convex sets. b. Gappy sets.



lect MMEU and then dial down the pessimism with a suitable choice of cautiousness function, or we can select some more ambiguity loving rule and then dial up the pessimism. How are we to know what cautiousness function is to be added to MMEU in order to achieve the ambiguity level represented by, say, $\alpha = 0.7$? I do not know the answer to this, and I recognise that it is an important question for implementing the confidence approach. More work needs to be done on measuring ambiguity attitudes, and connecting those measures with this framework.

Stakes dependence is not the same issue, however. It *does* create room for optimism or pessimism of a sort, but this should not be confused with ambiguity seeking or ambiguity aversion. An agent who ranks the importance of decisions on the basis of the best possible outcome only is certainly more optimistic than an agent whose importance ordering is formed on the basis of the worst possible outcome. But this is not a function of that agent's uncertainty about the states of the world, nor of their attitude to that uncertainty. Furthermore, how "optimistic" an agent is in their stakes evaluation is a feature of that agent, and is not subject to the same representational flexibility that is present with ambiguity attitude. Indeed, as Hill puts it, "different notions of stakes will correspond to different behavioural properties, and hence it is possible in principle to tell whether the decision maker is using a given notion or not" (Hill, 2016, p. 86).

6

CONCLUSION

This concludes my philosophical study of policy-making under scientific uncertainty. I considered two manifestations of uncertainty—expert disagreement and model ensembles—and have made a tour through various approaches to them, all falling broadly under the umbrella of formal epistemology and decision theory. My goal has been to find or create tools that are fit for the purpose of guiding or supporting real policymakers making decisions in the face of these forms of uncertainty.

In the introduction I characterised my project as non-ideal theory, or at least less-ideal theory. This has informed the route that my tour has taken.

In chapter 2 I examined broadly Bayesian approaches to expert testimony and expert disagreement. The chief contribution of that chapter is the development of a new model of expert deference, designed to solve a set of problems with two more orthodox approaches: supra-Bayesianism and expert deference as a constraint on priors.

I argued that supra-Bayesianism was unreasonably demanding, that it made critical and worrying use of the policymaker's uninformed priors, that it was insufficiently sensitive to the testimony of the experts, and that it assumed that the policymaker is always antecedently aware of the propositions the expert reports on. I argued that expert deference as a constraint on priors improves on some of these issues, but introduces two additional worries. The first is that it arbitrarily isolates certain reports—reports of unconditional probabilities—as alone being worthy of deference. The second problem is that it gives us no grip on the problem of expert disagreement.

I developed a proposal for regarding expert deference as a belief revision schema. In my model the expert reports are externally given constraints on the agent's posterior beliefs. The Bayesian updating procedure is replaced with a two-part belief revision process: expert deference (now realised as the imposition of this external constraint) and a suite of belief revision rules that "complete" the posterior. Each rule is matched with a type of input for which it is appropriate.

I argued that this new model improves on both supra-Bayesianism and the orthodox model of expert deference, making progress on all but one of the issues highlighted above. However, the remaining issue—expert disagreement—is also the focus of my study. So this first leg of the tour ends with mixed success.

We have seen how Bayesian ideas and tools can be adapted to suit more limited agents. My new proposal has as a particular benefit that it can accommodate a wide range of kinds of expert report. It also shows promise for integrating with solutions to the problem of rational unawareness and rational awareness growth— two key issues for the study of boundedly rational agents.

Although my new model makes limited advances on the issue of expert disagreement, there are exciting prospects for future work here. In particular, the problem of modelling awareness growth could be treated in much greater depth. Studying awareness growth in the context of expert testimony, as I've done here, illuminates something about the structure of the problem.

Experts only report on so much; so the agent only learns the probabilities for the propositions reported. But, as my example showed, even relatively simple cases of awareness growth can involve the introduction of a great many new atomic propositions, at a finer level of grain than that reported on. Unless the agent is given sufficient information to fix the probabilities on these new "worlds", their posterior credence function will be radically under-determined. Building on this work, one could look for results characterising the under-determination of the posterior, given the nature of the awareness growth experience.

Incorporating additional tools from lattice theory should allow one to make precise the sense in which awareness growth is the "reverse" of Bayesian conditionalisation, thereby making precise the intuition behind the name Karni and Vierø (2013) gave to their proposal. While conditionalisation takes one from an initial algebra to a sub-algebra, awareness growth takes one from a initial algebra to a super-algebra. Bayesian conditioning is one way of forming a new probability function on the sub-algebra after learning. Bradley's rigid extension is the natural inverse procedure for extending a probability function to a super-algebra. One chooses a probability on the super-algebra such that, if one were to condition on the initial algebra (now a sub-algebra) one recovers something like the prior.

In chapter 3 I turned to a study of the set of averaging methods known as opinion pooling. I was concerned with whether we should average opinions; if so, how we should do it; and in particular, how we should weigh the contributions of the different experts. Four important conclusions emerged.

First, I found some reason, albeit limited, to use linear averaging. There is a decent body of evidence that shows that average forecasts often outperform individual experts over time. This gives us a prima facie reason to be interested in linear averaging but does not tell us how or when it works. A mathematical result shows that we should choose the linear average, when we are in a particular situation: facing a choice between a randomly selected expert and the average, where the success of our chosen expert will be measured by a convex measure of error. In this circumstance, the average will always weakly outperform a randomly selected expert.

Second, I also argued that in cases like mine, where we are concerned with actual policy decisions that make use of expert panels, we should not be overly concerned with the so-called rationality axioms governing the choice of pooling function. External Bayesianity and Individualwise Bayesianity aspire to a rational ideal that is unjustified, and are aimed at avoiding a form of manipulation that is simply unlikely to arise in expert elicitations.

Third, I argued that under certain assumptions, the science of scoring rules plays a peculiarly important social epistemic role as the arbiter of expert disagreements. I noted the existence of a plethora of scoring rules, and observed that the technical choice between these is inaccessible to the policymaker. The disagreement of philosophical and statistical experts about which rule is best presents them with a second expert disagreement; threatening to undermine the usefulness of averaging as a solution to (the policymaker's initial) expert disagreement. I argued, however, that if opinion pooling is justified, then this is not a reason for despair, but rather the identification of an important form of meta-expertise. If disagreements amongst various scientific experts are to be resolved using averaging, then policymakers would do well to invest in the expertise required to successfully conduct expert elicitations. Each of us faces the choice of how to invest our limited cognitive resources, and for the policymaker I argued that an investment in the science of scoring probabilistic predictions seems likely to yield high returns.

Fourth, I highlighted the often hidden role that non-epistemic values play in selecting a scoring rule. This highlights the importance of policymakers playing an active part in any pooling procedure, and reinforces my proposal above: that policymakers learn the details of scoring probabilistic predictions, so that they can effectively exercise their values (or those that they represent).

Chapter 4 extended this analysis to the case of model ensembles. In particular, I introduced my case study on models of North Atlantic hurricane formation and development, and argued that averaging the results of this ensemble is unjustified and likely to lead to poor decisions. This is a direct criticism of the current practice in the hurricane insurance industry.

I argued that many of the justifications for averaging model results fail under scrutiny. Despite the wide popularity of statistical methods for treating model results, they are not like measurements and the collection of model results is not a sample. The use of statistical methods for treating samples is not well justified in the case of model ensembles. Similarly, model results are not plausibly subject to a convergence result, such as the Condorcet jury theorem. Models are not independent, nor do we have good reason to believe that they are minimally competent. (It is also unclear exactly how this analogy is meant to work, but the previous problems are more important.)

The problems of scoring rule choice, discussed above for expert opinions, apply to the case of model averaging too. Selecting a rule on purely technical grounds would result in some degree of arbitrariness. Different rules, each roughly equal in terms of technical suitability, may lead to quite different averages. Any such selection will make implicit value commitments that are properly the domain of the policymaker. (This difficulty could be mitigated by enacting the changes described above, but in the currently prevalent procedures it is present.)

Even if averaging were justified, it requires more data than is available in the case of the hurricane model ensemble. Crucially, averaging often obscures the actual level of scientific uncertainty. It focuses attention on the mean probability, and does not make the decision-maker aware of the spread of possibilities that experts thought were reasonable. This is in part a contingent fact about decision procedures that make use of averaging: it is often regarded as a technical process that precedes the communication of results to the decision-maker. But it is also a problem with the dominant decision theory: there is simply no role for information about the spread of model results in an expected utility decision framework (no role, that is, over and above its role in determining the average probability function). Finally, I highlighted particular cases in which averaging yields counter-intuitive results, by violating agreement over what the right answer isn't.

This laid the ground for a new approach to deciding using model ensembles.

In chapter 5 I presented such an approach: the ambiguity decision theory developed by Brian Hill (2013, 2016, 2019), that I call "the confidence approach." I argued that the confidence approach improves on the issues just highlighted for decision-making using the average probability.

Most of chapter 5 is a demonstration of how the confidence approach handles decisions with inputs from model ensembles. As we currently lack good tools for making decisions of this sort, my core task was demonstrating its suitability and illustrating its virtues. I hope that this motivates policymakers to investigate the approach as a decision-support tool, and that it motivates philosophers to further study and refine the confidence approach.

The chief benefits of the confidence approach are these. It is as explicit as possible about uncertainty at every stage, making use of it rather than obscuring it. Where it makes use of a skill score, the approach mitigates against worries over its selection by building in a flexible degree of robustness. Decision-makers are presented with a variety of options: different sets of probabilities, each with an attached "cost" to their use in the form of the confidence it can support. The approach thus builds in a natural link between the importance of the decision, the confidence that importance demands, and the formulation of decision-input.

For hurricane insurance (the immediate context of my study) there are some more particular benefits. The current uncertainty management procedure involves averaging model outputs and making ad hoc adjustments based on underwriters' judgements. This is unsatisfactory for various reasons, not least that it is unsystematic and opaque.

The confidence approach allows insurers to approach the problem systematically. The approach is suitable for implementation in a firm with distributed decision-making, as the ingredients (stakes, cautiousness, decision rule) can be isolated and set as a matter of policy. It provides a recipe for pricing insurance which is sensitive to all of the evidence available for that decision, and which responds naturally (through the cautiousness and stakes) to the different nature of each decision taken. This kind of flexible but systematic treatment of uncertainty is what insurers tell us they have been missing in catastrophe insurance.

The confidence approach is my preferred solution to the problem of policy-making under scientific uncertainty. Like the other solutions that I considered, it has some outstanding issues and there is room for further work. I believe that it improves on the methods discussed in chapters 2–4.

I have extensively discussed how the confidence approach compares with averaging, but it is worth noting that it also improves on the Bayesian methods discussed in chapter 2. In the form presented here, the confidence approach does not make worrying use of the policymaker's uninformed priors. Instead, all the probabilities involved are supplied by the relevant experts or models. It is therefore sensitive to the testimony of experts. It is expressly developed here as a method for dealing with expert disagreement. It also avoids the problem of awareness, though in a somewhat limited manner: by having no aspiration to be a complete theory of the policymaker's doxastic life.

There are significant avenues for further work. First, there are avenues for further development of the decision theory itself. The way that the confidence approach represents ambiguity attitude is worthy of deeper study. Some of this work will involve refining the theoretical apparatus to more clearly articulate the attitudes involved and how the machinery represents them. The literature on measuring ambiguity attitudes is large, and may contain tools that can be deployed in resolving how to represent that attitude in this approach.

Second, the link with model ensembles can be developed. The material presented in section 5.5 is only a preliminary investigation into different methods for constructing confidence rankings using model ensembles. A study of the detailts of particular models and their evidential bases could fill in the details left schematic in chapter 5.

A natural extension would be the application of confidence methods to models of the climate system. In that field it is already commonplace to speak of probability assessments as coming with different degrees of confidence (e.g., Winsberg, 2018, Ch. 7). This forms the basis of some work linking the confidence approach with the IPCC's uncertainty framework (Bradley, Helgeson, and Hill, 2017, Helgeson, Bradley, and Hill, 2018), but other connections remain unexplored. There is a large literature on uncertainty in climate science, and on the usefulness of ensemble methods in managing that uncertainty. By providing a new decision tool suited to taking inputs from ensembles, the confidence approach may provide resources to advance that discussion.

In this thesis, I hope to have provided some insights and tools into an exceedingly difficult class of problems. The fact that there is such a rich set of future research directions is, I think, a sign of the value in approaching the problem as I have: by modifying the powerful tools of formal epistemology and decision theory so that they are suitable for bounded agents working with severe uncertainty.

7

CODA: FORMAL PHILOSOPHY AS MODELLING

7.1 INTRODUCTION

This chapter is a coda to this thesis, reflecting on its methodology. In writing the preceding chapters, I had the opportunity to work in two often disconnected fields: decision theory/formal epistemology and the philosophy of scientific models. While the last two chapters involve applying tools from the former to a problem in the latter, this chapter does the reverse. Here I turn a philosophy of modelling eye on formal epistemology and decision theory.

This chapter has two aims, beyond self-reflection. First, talk about "modelling" as a method of philosophical inquiry is increasingly prevalent and in need of explanation. Williamson (2006, 2017) names modelling as an important method in a certain style of philosophy (what we might call scientific or mathematical philosophy). He defends modelling as an important tool for developing clear arguments (2006, pp. 186-7), and as a major source of philosophical progress (2017, p. 8). Stephan Hartmann often uses and discusses models in philosophy (e.g., Bovens and Hartmann, 2003; Eva and Hartmann, 2019). Hannes Leitgeb (2013, p. 273) mentions modelling as a method of building inductive strength in an argument. Peter Godfrey-Smith (2006, 2012) and L.A. Paul (2012) discuss modelling as a practice in metaphysics. Michael Titelbaum (2012) describes the project of his book as providing a framework for building models in formal epistemology.

In all these cases, the talk of modelling and model-building is an analogy with the commonplace scientific practice of indirect inquiry using idealised representations. This is part of a wider naturalistic turn in thinking about philosophical methodology. I want to advance this discussion by providing a clear explanation of how the scientific methodology of modelling can work in philosophy. There are crucial differences between the philosophical and scientific cases that need to be examined to ensure that it is the same method. I will focus on one: philosophy is often normative, and the objects being called "models" in philosophy often serve normative purposes; while science is not typically normative, and models in science typically serve descriptive/explanatory/predictive ends.[1] How does this difference influence the claim that we are (sometimes) modelling in philosophy? I will argue that normative work can be considered modelling, though there are some unique considerations in the normative case concerning the role of idealisation.[2]

Though I will not comment on ethics directly, an account of normative modelling in philosophy will, I think, be of great use to moral and political philosophers. There is a long history of discussion about

[1] Note that our subject here is not representational models of communities obeying norms, or of how norms might emerge, such as those studied in the literature on the social evolution of morality. I am here interested in models whose purpose is the generation of normative claims.

[2] Colyvan (2013) has discussed the role of normative assumptions in formal models such as Bayesian decision theory and logic. My project differs in that I provide an account of the content of normative models, and their relation to the world, along with more explicit methodological conclusions for FE.

the role of idealisation and abstraction in ethical theory, (e.g., O'Neill, 1987), and how it relates to the distinction between ideal and non-ideal theory (e.g., Mills, 2005). As some have noted (e.g., Hancox-Li, 2017) there are obvious parallels between idealised, abstracted ethical theorising and modelling. My account of normative modelling and its role in another normative field (epistemology) will hopefully be of use to that discussion, though additional work will be required to reap the benefits I promise, as my focus here will be on *formal* normative modelling and much of the relevant ethical work is not formal.

Second, the growth of formal epistemology (FE) has led to a spread of its ideas into more mainstream philosophy without a corresponding dissemination of its methods. Many philosophers now consider the attitude of partial belief (or degree of belief) to be an important topic in epistemology; typically in the form of "credence", a particular mathematical representation of that attitude as a number between 0 and 1. This chapter aims to develop an explanation—for a wide philosophical audience—of what formal philosophers are up to, and how one should regard objects like credence. My conclusions are cautioning: normative work using models is complex, and a number of inference-patterns familiar from other parts of philosophy do not work well here, including certain realist inferences and reasoning by counterexample.

In section 7.2, I will describe my target more completely by outlining one example of what formal epistemologists call a model. In section 7.3, I present a loose characterisation of scientific modelling, and some lessons from the philosophy of science literature on it. These are deployed in section 7.4 to characterise formal epistemology as modelling in a first pass, and then extended in section 7.5, where I present my account of normative modelling. In section 7.6 I consider alternative explanations of what formal epistemologists could be doing—modelling is just one methodology among many, and it is helpful to distinguish between them, both to delineate the alternatives and to highlight the contrast between my account and the commonplace view of philosophical methodology. I then turn in section 7.7 to some methodological considerations for FE, given that we are modelling. A key part of the discussion is a consideration of when normative conclusions drawn from models are "secure"—well-justified by the model—given the dependence of modelling on idealisation. Section 7.8 concludes.

7.2 THE TARGET

Though my conclusions apply broadly to modelling in formal epistemology, decision theory, formal value theory and perhaps beyond, I will work primarily in the context of formal epistemology (FE) for clarity.

Let us start with an example of the kind of structure that I want to call a model in FE. Consider some common modes of inquiry in formal epistemology, concerning partial belief.

- *The nature of rational partial belief.* Here, we translate norms of rationality into a formal (i.e., mathematical) setting, and use the precision this affords to draw conclusions about the implications of those norms for the structure of our attitude of partial belief.

- *The norms of rationality.* In another mode, the norms themselves are at issue. We examine the plausibility of putative norms by translating them into a formal setting, deriving mathematical results, translating these results back into ordinary language, and testing them against firmly-held intuitions.
- *Decision-making.* Pairing norms of rational belief with norms of rational desire allows us to derive rules for selecting one option from a menu of possible acts. Again, we might explore the implications of accepted norms or test the implications of putative norms.

In each case, we start with an initial question or problem about the attitude, framed in natural language. Some principles of rationality are chosen to govern the agents involved. These are translated into a formal language capable of representing agents, propositions they consider, beliefs they hold, and so on (as necessary). Constructing this formal apparatus typically requires introducing additional structure, that is not motivated by the initial question but is internal to the process of representing it mathematically. The formal setup is then studied, and conclusions are drawn. Finally, these formal results are translated into conclusions about partial belief or decision-making.

Here is a specific example. We might begin with some observations about our subjects: people have partial beliefs. (Partial beliefs are often communicated in the language of likelihood and referred to as "comparative likelihood judgements", or "comparative confidences".)

We observe people making statements about their confidence in various judgements or making comparative judgements about two propositions they avow to believe, such as "I am fairly certain Brexit will be a disaster" or "I am more confident that it will rain tomorrow than I am that Boris Johnson will make a good Prime Minister." Under good conditions, these partial beliefs seem to have the following properties:

- "Monotonic": we believe weaker propositions to a greater degree than stronger. I believe "it will rain on Monday or Tuesday" more than I believe "it will rain on Monday".
- "Separating": we can "factor out" common propositions when making comparisons. If I regard it as more likely that it will rain on Monday than on Tuesday, then I regard it as more likely that it will rain on Monday or Wednesday than on Tuesday or Wednesday.
- "Transitive": If I believe it is more likely to rain on Monday than Tuesday, and more likely on Tuesday than Wednesday, then I must believe it is more likely to rain on Monday than Wednesday.

In the first two cases, "under good conditions" means something like "when we're aware of, and think consciously about, the logical relations between the relevant propositions." In the third case, it means something like "when the initial two pairs of comparisons are considered together." These patterns strike us as reflecting something about the logic of partial belief, and as we begin to theorise this attitude it seems that believing in a way that doesn't fit these properties would be doing something wrong. These therefore become putative norms for rational partial belief.

This is our topic of study. What distinguishes formal epistemology is the decision to represent this attitude with a mathematical object. In this case, it is common to represent partial belief with a binary relation, which I will denote by $\succeq$, that encodes an agent's comparative judgements. Writing R for "it will rain tomorrow" and B for "Boris Johnson will make a good Prime Minister", $R \succeq B$ represents the judgement that I am more confident that it will rain tomorrow than I am that Boris Johnson will make a good Prime Minister. I will refer to $\succeq$ as the "comparative credibility" relation, in order to avoid re-using the term confidence. Credibility is defined on a Boolean algebra Ω, a set of propositions which is closed under $\models$, an implication relation. This implication relation can be defined as $X \models Y \iff X \vee Y \models Y \iff X \wedge Y \models X$, for classical con/disjunction (Bradley, 2017). We endow credibility with mathematical properties that correspond to the attitudinal properties we settled on above: it is monotonic, $\vee$-separable (Joyce (1999, p. 91) calls this "quasi-additive"), and transitive. We take these mathematical properties to represent the norms that we theorised for the attitude of partial belief, just as credibility represents that attitude.

Binary relations are not particularly easy to work with, and so formal epistemologists typically make progress by using a "representation theorem". A representation theorem is a mathematical argument showing that a binary relation $\succeq$ can be represented by a real-valued function $F : \Omega \to \Re$, where this means just that $F(X) \geq F(Y) \iff X \succeq Y$. Such a theorem typically specifies the form of the function, and some uniqueness conditions for it. Under certain conditions, credibility can be represented by a probability measure. I will use common jargon and call this a "credence" function, denoted P. For credibility to be represented by a credence function, it needs to have certain mathematical properties. The details vary depending on the particular representation theorem, but if we examine important theorems which result in unique probability functions—Villegas's theorem and Joyce's theorem—we observe two kinds of conditions. The first are precisely those normative conditions discussed above: monotonicity, transitivity and separability. The second includes, for example, the requirement that credibility must be *complete*: for any two propositions $X, Y \in \Omega$, they must be related in some way by $\succeq$: either $X \succeq Y$, $Y \succeq X$, or both.[3]

We will return to the details of these representation theorems below, but for now we note that completeness is not a particularly compelling norm for partial belief. It is therefore common to regard this second group of requirements on credibility as *non-normative*. The model therefore comes to include both normative and non-normative assumptions about its mathematical consituents.

The theory that has just been formalised is intended to be normative, however. So, from the fact that partial beliefs can be represented by probability functions (under certain conditions), we derive a norm

[3] More fully: these two theorems require credibility to be monotonic, separable, transitive, complete and continuous. The Boolean algebra must also be complete and atomless. Bradley (2017) discusses each assumption.

for partial belief: if your partial beliefs cannot be so represented, then you are irrational. (This norm is called Probabilism.[4])

Two comments are important at this early stage about this normativity. First, the primary mode of normativity operant here is that of evaluation. This is a standard against which we are measured; it is out of reach, but linked in important ways to our actual capacities. There is a secondary mode of normativity—prescription, or action-guidance—that is largely present in some parts of decision theory. I will focus on the evaluative mode here.

Second, there are two common ways that such formal work takes place in epistemology. We might work constructively, introducing and defending each assumption about credibility in turn and concluding with the norm of Probabilism. The defences are typically that the assumptions are themselves norms (e.g., monotonicity as I presented it above), or that they are true descriptions of the attitude, or that they are harmless structural requirements—mathematical conditions that don't represent anything but are useful to get the discussion moving.[5]

Alternatively, we might work critically, by starting with a formal apparatus and criticising it for making the wrong ruling: either it declares something to be bad that is in fact good, or vice versa. (I'm using "good" and "bad" here for the two valences of the relevant norm, e.g., rational and irrational.) In decision theory, these critical engagements often involve particular choice situations, like the Allais

[4] There are, of course, other (better) ways to argue for this norm. I don't claim this is how it ought to be done, but grant me that it is sometimes defended in this way for the purpose of the discussion.

[5] Not everyone is so cavalier, of course! Joyce thinks "that the Achilles heel of Savage's theory is its dependence on structure axioms that cannot be satisfactorily explained away" (Joyce, 1999, p. 98) —a conclusion quite close to mine in this chapter, although not presented in anything like the same way.

and Ellsberg “paradoxes”. A particular decision theory rules the Allais choices irrational but, says the critic, they are intuitively rational, and so that decision theory is flawed.

Some of the above is obviously similar to scientific modelling—the use of mathematics in a representational role, the presence of idealisations. But some is peculiar to philosophy. So, are these models? If so, how do they work, and does their methodology come with the same constraints and benefits as scientific modelling?

7.3 SCIENTIFIC MODELS

Let us begin with a review of scientific modelling, and the methodological lessons we have learned from five or six decades of philosophical study of modelling.

“Model” is one of those unhelpful terms that is used to mean many different things, so I want to begin with a common meaning that I *do not* use: the meaning logicians give to the term. Roughly put, logicians use “model” to mean an interpretation that satisfies a set of sentences. An interpretation is here an assignment of semantic values to the basic vocabulary in use. This semantic sense of “model” takes it to pick out certain *mathematical structures*. Some philosophers of science (e.g., Suppes, 1969) have argued that this meaning of “model” is the same as, or should be used to explicate, the workaday use of “model” in scientific practice. This is a view which is associated with the once popular “semantic view” of a different scientific construct: the theory. I will not be using “model” in this sense, and in that I will diverge at the outset from some (like Paul, 2012) who have discussed modelling in philosophy. The way I use the term "model" is broadly consistent with a philosophy of science tradition that includes Giere (1988, 2004) and Cartwright (1989) as well as the many others cited below, and is more or less how Godfrey-Smith (2006) uses the term.[6] If you typically think of models as set-theoretic structures you will need to take this section as stipulating a new meaning for that term.

So what is a model? Here are three examples, to ground intuitions as I introduce the theoretical account. Some models are material objects, like the molecular structure models used by chemistry students. Modelling kits, such as the MolyMod system, come with coloured balls representing elements (white for Hydrogen, red for Oxygen), and grey connecting rods representing chemical bonds (short and stiff for single-valence, long and bendy for double-valence). With these kits, students build models of simple molecules like H_2O, and more complex polymers like PVC. We call the real-world system under study the target, and the plastic object the model. The model of H_2O involves one large red ball connected by two short grey rods to two smaller white balls, in a wide V shape. The student learns about the structure of the molecule, H_2O, by examining the plastic model.

More commonly, models are theoretical rather than physical. The Bohr Model of the atom is a classic example: Bohr imagined the Hydrogen atom as an orbital system consisting of a central positively charged sphere orbited by a distant, negatively-charged sphere. The

[6] I will not attempt a classification of all of those writers about philosophical modelling mentioned in the introduction. For one thing, many writers switch between different senses of the term model in the same paper. This is the case with Paul (2012), who appeals to the work of Godfrey-Smith (2007) alongside semantic-view authors whose views Godfrey-Smith explicitly repudiates.

centre represents the nucleus, the orbiting sphere represents the electron. The "electron" is in a circular orbit, and only certain orbits (with specific orbital distances) are allowed. What is the model in this case? It is described by a series of written statements (like those above, together with Bohr's "rules" for electron orbits), often accompanied by an equation (e.g., $L = n\hbar$, specifying the angular momentum of the orbiting electron) and perhaps illustrated with a diagram. But the system we are investigating when we use the Bohr Model is not identical with any, or all, of these physically instantiated parts; it is what those descriptive elements specify (Mäki, 2009, p. 33). There are several philosophical accounts of what such model systems are, but for now we need only note that, whatever they are, they are non-physical.

Finally, some models have no target in the real world. Architectural plans for buildings which will never be built are models, as are theoretical models for ether or phlogiston, substances which do not exist. In modern quantum field theory, "ϕ^4 theory" is a simple, intuitive model which has been extensively studied despite being known not to correspond to any physical system (Frigg and Hartmann, 2018). So, a philosophical account of models must rely neither on a concrete model system, nor on a concrete target system.[7]

Philosophers of science have developed a rich literature on the representational function of models, their ontology, epistemology, and implications for scientific realism (see Frigg and Hartmann, 2018). I

[7] I call these "target-less models". Some, notably Weisberg (2013) distinguish between models where a target system is discussed but it does not exist (like an architectural model) from models where there is no specified target, real or imagined, like the "Game of Life", a cellular automaton. I am not convinced that the Game of Life is a model, rather than a piece of interesting mathematics that has inspired thinking relevant to various sciences. I won't argue for this here, but it is the reason that I prefer to eliminate the distinction between hypothetical and target-less models.

will here draw attention to a few lessons learned in this literature, for comparison with the practice of formal epistemology. [8]

(1) Modelling is characterised by indirect inquiry (Giere, 2004; Godfrey-Smith, 2007; Weisberg, 2007b). Instead of studying the natural system, modellers describe and investigate a "model system" which is the primary target of their investigation. The model system is taken to (partially) represent the target natural system. Modellers then infer facts or generate hypotheses about the target system based on their investigation of the model system. (In cases where the model is target-less, they are still thought to be representational in a sense to be discussed later.)

(2) Models present an idealised and distorted picture of the target system (Frigg and Hartmann, 2018; Weisberg, 2007a). Many real-world systems cannot be investigated directly, due to incomplete theories or severe computational complexity. To make progress, scientists simplify the system under investigation, by changing or leaving out aspects of the real system. They work to identify the features of the system most salient to their investigation (Weisberg, 2013, p. 4). The frictionless plane is a classic example: no real surface is frictionless, but it is fruitful to take a surface to be frictionless when investigating the inertial motion of objects on an inclined plane.

There is an extensive literature on idealisation in science; I will note two distinctions drawn in that literature for use here. There are different kinds of idealisations: Galilean and Aristotelian (Frigg and Hart-

[8] While they are not without opposition, I aim to use only "mainstream" views in the philosophy of modelling. I also do not attempt to provide anything like a complete bibliography on each point here. Rather I cite a few recent sources, with good references in each for the interested reader.

mann, 2018). Galilean idealisations introduce deliberate distortions to some properties of the system under investigation. For example, the friction of the plane is deliberately changed in the representation. Aristotelian idealisations leave out features of the system that are not relevant to the problem being studied, to allow us to focus on or isolate a limited set of properties. For example, a population growth model considers only the rate of reproduction and predation of organisms and leaves out other properties such as their physical size, colour, and social structure.

There are also different motivations for idealisations (Musgrave, 1981). A modeller might take a property to be negligible, believing that for the purposes of the current investigation it will make no difference to distort or exclude it. For example, we might consider falling objects, and idealise by assuming there is no air resistance because we believe it to be of negligible importance. Another way of putting this is that the idealisation functions well when it is true that the effect of air resistance is small, so that the model's claim that air resistance is zero is approximately true.[9] Alternatively, the modeller might know that the property is not negligible in all cases but want to model only those cases where it is so. Musgrave calls this a domain idealisation: it justifies itself "automatically" by restricting the class of cases the model applies to. Finally, the modeller might think that there are no cases where the property is negligible but distort/exclude it anyway because its presence in the model makes things too complex to handle. Musgrave calls this a heuristic idealisation, and

[9] I don't want to be committed to an approximate truth account of idealisation here; I am merely presenting some ways idealisations are thought of.

presents it as part of a process of inquiry: we simplify the model by setting air resistance to zero now, with the hope that once we have established the model we can factor air resistance in later. Note that negligibility, domain-restriction and heuristic necessity are species of justification—the same idealisation can be justified in each way, depending on the modeller and the circumstances.

(3) Models are built for a purpose, and so perform well only within a restricted domain of applicability (Parker, 2009a; Teller, 2001; Weisberg, 2007b). "Purpose" consists of what you're modelling (e.g., ants rather than bears) and what you're trying to do (e.g., study group coordination). This establishes the basic domain of the model (it is a model of ant coordination). As Wimsatt (2007, p. 15) points out, models are often used to isolate particular mechanisms or concepts for study. This purpose motivates the idealising assumptions, which may further restrict the domain of applicability as discussed above. I'll refer to the combination of purpose and domain as the model's scope.

The purpose-driven nature of modelling means that model-based sciences often contain multiple, disagreeing models of the same phenomena. Teller illustrates this with an example of two models of water. The first is interested in the flow of water and wave propagation, and it models the liquid as a continuous incompressible medium. The second is interested in explaining diffusion, say of a drop of ink in water. It models water as a collection of discrete particles in thermal motion. Each is similar to water in the respects that are relevant to its purpose, but the two models look very different (Teller, 2001, p. 401).

Neither should be thought to provide a definite characterisation of water, and our understanding of water is enhanced by having both available.

7.4 THE METHODOLOGY OF MODELLING

The foregoing characteristics of modelling and models lead to certain methodological constraints for this kind of science. Idealisation is the lifeblood of modelling, but while it helps scientists make progress in investigations of complex systems, it introduces limitations. As Levins (1966) put it, modelling involves an inherent three-way trade-off between precision, realism and generality of scope.

On the realism front: models contain artefacts, properties of the model system that are not representative of any real feature of the target system, but instead emerge from the representational choices of the modeller or the idealisations in the model. Good modellers must identify artefacts and ensure that they aren't imputed to the target. If there is an underlying fundamental theory (as if often the case in physics), this can help to identify artefacts. Another method for identifying such effects is sensitivity analysis.[10] This is a method for studying the uncertainty of a model, and allocating it to the sources of uncertainty in its inputs. In the use I am considering here, it involves varying assumptions in order to determine the effect that these variations have on the results. For example, let us consider again an idealisation of no air resistance, justified by a negligibility assumption.

[10] Also called stability analysis, it is closely related to what Weisberg calls "parameter robustness" (Weisberg, 2013, p. 159).

If we have set the parameter representing air resistance in our model to $k = 0$, we might vary this by considering small but non-zero values of k (small relative to some natural scale determined by the problem). The aim is to ensure that the results we get don't depend sensitively on the air resistance being exactly zero, and simultaneously to test that the negligibility assumption (about the real system) holds in our model—i.e., that small values of k make only small changes to the results.

The result of this kind of investigation is what Frigg and Nguyen (2016) call a "key". By analogy with a map's key, this is a legend that tells the user how to interpret what they're seeing. It specifies how results from the model should be taken to relate to the world, covering issues of realism and precision: a key might specify that some precise number generated by the model should be taken as a prediction for the real system only to within some error-margin; or it might identify some element of the model as an artefact, not to be imputed to the target at all.

This trade-off is thought to prevent theorists from developing a single "best" model for a complex system (Levins, 1966; Weisberg, 2013, Ch. 9). The resulting prevalence of multiple models of a single system also has methodological implications—most straightforwardly, we cannot take disagreements between, e.g., Teller's two models of water, as a sign that one of them must be rejected. Each can be useful for its purpose. Wimsatt (2007, p. 104) highlights that multiple idealised models can support the development of fuller theories, through the examination of results on which all models agree. This is a particularly useful technique in situations without underlying fundamental theory, such as some areas of biology (Weisberg, 2013, p. 156). This is called "robustness analysis" and its aim is to find robust results, or "robustness theorems".

As the above implies, criticising models is a complex business. As models have restricted domains, and specific purposes, the most natural way to critique a model is by examining how well it performs its purpose within its domain. Performing poorly on other tasks, or in other domains, does not count against a model. It can do so if two models are being compared, and the one performs better on the shared purpose, and has wider scope (either wider domain or the ability to fulfil multiple purposes). Put another way, models are not sensitive to counterexamples the way that fully general accounts are. Saying "here is a case that isn't like your model predicts" matters only if the case is in scope. Similarly, saying "your model says things are like so, but here is a case where they aren't" only matters if that feature of the model is intended to be imputed to the target. If the model's key identifies the feature as an artefact or says it should be imputed in some modified form, then the disagreement between the model's properties and the target's properties is irrelevant.

7.5 NORMATIVE MODELS

Having reviewed these lessons from the philosophy of scientific modelling, I now turn to our main topic: normative models. My aim is to argue that formal epistemology fits the characteristics that define modelling, and therefore that the methodological considerations discussed above apply to FE too. But in order to do so, and indeed in order for model-talk to go through, an important task remains. We need to provide an account of the main difference between inquiry in FE and the sciences: much of FE is normative. Our efforts are directed at what agents ought to do, rather than at explaining or predicting what they do. This section develops such an account. In addition to explaining what it means to say that something is a normative model applicable to real agents, I also want to vindicate our other common way of speaking: describing FE models as models of ideal rational agents.

To begin, however, let us note that normativity is not so foreign to science. Physiological models in medicine can be thought of as normative, representing how the body should be, with real deviations representing illness. Economic models of perfect competition might be taken by economists to specify how a market should work, with deviations representing barriers to be overcome, "imperfections" to be removed. Ecological models might represent an undisturbed ecosystem and thereby act as an evaluative standard for assessing the impact of alien species. Social choice models of voting procedures act as blueprints for the design of real voting mechanisms. Architectural models describe how buildings ought to be built.

Some of these involve a weaker sense of "normativity" than that familiar from ethics. But so long as normativity is understood as meaning "subject to judgements concerning oughts" then we can happily describe the above as normative for some sense of "ought". Nonetheless, current work in the philosophy of scientific models does not focus on these normative aspects, preferring representation as a topic of philosophical discussion. My purpose in highlighting these models at the start of this section is to point out that addressing normativity is not a concern peculiar to the application of model-talk to philosophy. Indeed, I will build my account of philosophical modelling by first considering models from outside of philosophy that play normative roles.

7.5.1 *The architectural model*

I will start with an example of one such normative model in science, as a guide to our thinking. Consider an architectural model of a block of flats. When architects first develop such a model, it serves mostly as a vehicle to communicate design ideas. The drawings are typically rough and impressionistic, representing high-level aesthetic ideas and establishing basic features of the building such as floorplan layout. As the building project advances, the model shifts to a more exploratory mode. Constraints from physics and engineering are incorporated, and a more familiar scientific use becomes dominant. Architects use the model to examine the implications of putting a staircase here, or opening that floor to create more volume. Throughout the "design phase", it is a target-less model: it is not a representation of any existing building.

Later, the same model (now described by many complex drawings, outlining not just design and structural elements, but also services like plumbing and electricity) takes on a normative role for the construction team: it shows how the building ought to be built. There is a shift of audience over the design process, from client at the start to construction team at the end. For this latter audience, the targetless model becomes an instruction set for bringing a target into existence. There are two normative modes operant here: the model is an evaluative standard for the construction team's work, and it is action-guiding—skilled builders know how to translate the drawings into instructions. They build so as to bring into existence a building with properties as close as possible to those exemplified by this model. Once the building is complete the model becomes a familiar descriptive model, a representation of the new building (inevitably, an imperfect one due to deviations from the plan during construction).

This example gives us a handle on how normative modelling works. First, note that we have a movement back and forth between the model being descriptive and normative. This indicates that normative modelling is a way of using a model (rather than being a type of model). Which use it is put to depends on the purposes of the modeller/user and the intended audience. My first claim about normative models is thus: a normative model is any model that is put to normative purposes—evaluation, action-guidance, exploration of putative norms, and perhaps others.

Our philosophical models have similarly multifarious lives, a point that has been made in a different context by authors discussing the different "projects" of decision theory. Following Buchak (2013) we can distinguish four projects: construed normatively, decision theory can be used to evaluate or guide actions; construed explanatorily, it can be used to describe or interpret actions.[11]

The normative-evaluative use of decision theory involves analysing a decision situation facing an agent, and determining which actions are rational. The normative-action-guiding use involves deploying this process expressly in order to determine which act to undertake. The descriptive-explanatory use and the interpretive-explanatory uses are interested in real, rather than ideal, agents but they differ in their goals. Descriptive theorists are interested in describing observed patterns of behaviour—this is the empirical project of rational choice theory within economics. Interpretive theorists take real agents to be aiming at prescriptions of rationality but failing for various reasons. This theorist seeks to interpret the actions of the agent, as much as possible, as abiding by the rational theory of decision (this is often described as a "principle of charity").

The important point is that the *very same model* can be deployed in each of these projects (perhaps with different degrees of success). There need be no difference in the mathematical description of a decision theoretic model used in a normative-evaluative mode by philosophers interested in exploring the nature of rationality, and in a descriptive-explanatory mode by economists whose interest is in predicting choice behaviour.

Note that the sequence that occurred in the architectural case (first target-less representational use, then normative) is inessential: a model

[11] I have relabeled these projects for convenience, taking inspiration from Thoma (2019).

can begin life as normative, and later be taken up for descriptive purposes. The ability to switch between a normative and non-normative use of a model seems to be quite general. We can fix some actual system as a reference point for evaluation generating norms ("do as Buddha did"), and thereby turn a descriptive model of that system into a normative model; this secures the movement from representational to normative use. In the other direction: any normative model will presumably be relevant to some actual system (e.g., set of people) that lies within the scope of those norms. We can reinterpret any normative model as a (perhaps target-less) descriptive model of the relevant system where the norms are obeyed. Conceived of this way it isn't normative at all, it merely describes ideal agents, or a perfectly constructed building.

7.5.2 *Idealisation*

Our example "model" has normative constraints on the credibility relation, which correspond to what we take to be norms for partial belief. How can we fit these into the emerging account?

Above I argued that a "normative model" is simply a *use* of some model. Here, I will regard "normative assumptions" as a *kind of justification* for some idealisations. Recall that modellers might justify idealisations on the basis of negligibility, or because they delineate a domain, or as a heuristic device for making progress in early inquiry. Employing a normative justification for an idealisation is one way to put a model to normative use.

(My account comes with an implicit error-theory for some utterances of scientists and philosophers. I argue at various points in this chapter that philosophers *are* modelling, whatever they may say otherwise. Similarly, here I insist that some scientists—in this case economists—are wrong when they say that their models are "positive" despite incorporating rationality assumptions that those economists justify by saying that they are norms. If the reason your model says that preferences are transitive is that you think that they *ought to be*, then your model is normative.)

Consider our partial belief model, and the property of monotonicity. A positive economist might motivate this idealisation by any of Musgrave's three kinds of justification. The philosopher, by contrast, has a fourth option: whether or not it is approximately the case, or useful in simplifying analysis, partial belief ought to be monotonic over entailment. When philosophers and economists use "the same model," for normative and descriptive ends respectively, what they are doing is construing these conditions in different ways.

Our normative models also contain artefacts, which is to be expected given their use of idealisations. Consider logical omniscience, which is also exemplified by the agents in our model of rational partial belief. While it is hard to generalise about an entire discipline, my sense is that in practice logical omniscience is viewed as a mild embarrassment.[12] As I have set things up, it arises from monotonicity and the use of an objective logic to structure the Boolean algebra on which credibility is defined. These assumptions come before the representa-

[12] I report this as a sociological fact, to motivate for the existence of an implicit key. It is not universally true; some authors regard it as a serious problem—for a recent example see (Bradley, 2017, Part IV).

tion theorem that gives us credences, and one of them (monotonicity) was assumed to be a norm of rationality. As a normative demand, however, it seems excessively strong (indeed it violates a widely held intuition that ought implies can). But shifting away from using an objective logic is a daunting task—models of bounded rationality are often complex (e.g., Garber, 1984). Many philosophers simply mark this property as non-normative—we do not want to continually criticise agents for their lack of logical omniscience—and continue to use the model with that built into the key. The same goes for our agents' abilities of instantaneous computation. The fact that Bayesians disregard these properties is evidence that they are employing a key, which is characteristic of modelling.

The example of logical omniscience raises an interesting complication, which will be discussed further below. Some of the idealisations (such as monotonicity in my example) in a philosophical model will be normatively justified, while others (such as completeness) will be justified in one of Musgrave's three ways. But these idealisations can interact to create the properties of the model. What do we say about properties that depend on both normative and non-normative idealisations? Can a property be anything but an artefact, if it emerges only because of a heuristic assumption? I will return to this in section 7.7.

7.5.3 *Representation*

In what sense is the architectural model representational, given that there is no real-world system that it designates?[13]

Philosophers have answered this question by making use of Goodman (1976) and Elgin's (1983, 2010) notion of representation-as. The idea is to separate out two parts of our ordinary notion of representation. Consider the example of a famous caricature of Churchill as a bulldog standing on Britain. This is a representation of Churchill, but in a sense it is also a representation of a bulldog (after all, it is a picture of a bulldog with a vaguely Churchillian face). We separate out these two notions by calling the first kind, involving denotation of a target, representation-of; and the second, involving the way in which the target is shown (also known as its secondary subject), representation-as. The formula is: an object X (the drawing) represents a target Y (Churchill) as something, Z (a bulldog).

The Z variable identifies the secondary subject; the kind of representation it is, or what it portrays. We will refer to these genres of representation as Z-representations (e.g., bulldog-representations). One aim of introducing Z-representation is to show that there can be representation without reference to a target. A drawing of an orc is not a representation-of an orc, because there are no orcs. But the drawing does represent an orc in a sense, and we can now identify that sense by saying it is an orc-representation. As our formula says, Z-representations are objects (like drawings), and what fixes the

[13] This subsection follows Frigg and Nguyen (2016, pp. 226-28).

genre Z is an interpretation. We can think of an interpretation as a function, mapping properties of the object to properties of Z. We associate properties of the caricature (particular lines, shading etc.) with properties of a bulldog (a certain stoutness, folded skin, etc.).

Interpretations allow us to talk about Z-representations as "having" Z-properties that, strictly speaking, they do not. (The drawing does not have four legs, it is a drawing.) With something like a caricature, certain properties are highlighted as particularly relevant and the intention is that we impute those properties to the target. Bulldogs are pugnacious, and the caricature highlights this with the stance of the Churchill-dog in the drawing, with the intention that we regard Churchill as pugnacious. This is the final part of representation-as: when there is a target, we can impute highlighted Z-properties to the target.

A number of popular accounts of scientific models agree that they utilise representation-as (Elgin, 2009; Frigg and Nguyen, 2016; Hughes, 1997). A model consists of an object and an interpretation. The object can be concrete like the V-shaped collection of plastic balls and rods, or abstract like the mathematical objects of the Bohr Model. The interpretation specifies what kind of representation the object is intended to be, for example by connecting the balls and rods with elements and chemical bonds. The model object has certain properties: the colour of the balls, the structure of the ball-and-link system, the type of links present, etc. These are mapped by the interpretation to various molecule-properties, such as elemental composition and bond-structure.

With these elements we can describe much of the everyday practice of "modelling". Consider again the Bohr Model. The (theoretical) model object is the orbital system, specified by a set of descriptions and equations.[14] This is interpreted as being an atom-representation. The descriptions and equations are part of the theory of quantum mechanics, and they provide guidelines for the manipulation of the model—where "manipulation" would here involve calculation. Various results can be derived about the model-system, which are spoken of in the language of atoms due to the interpretation.

The "models are representations-as" account is of particular use in explaining target-less models, like our architectural model. We can now say that it is a building-representation that is not a representation-of any building (before it is built). The development of the building design involves manipulation of the model, to explore and communicate various building-properties. In its normative phase, it remains a building-representation and this representational aspect of the model is necessary for it to fulfil its normative role. The way that it serves as an evaluative standard for the building is by being a building-representation; exemplifying properties that the construction team's new building ought to have.

Our normative philosophical models have this representational aspect too. Our models of rationality are target-less; they are agent-representations that aren't intended to represent any real agents. These agent-representations are idealised, so the agents portrayed are dissimilar to real agents in various ways. Unlike scientific models how-

[14] I will not discuss what this theoretical object is; Frigg and Nguyen (2016, 2017) advocate for a form of fictionalism about models but there are other accounts available.

ever, some of these differences are regarded as normative. This account explains our language when we say that these are "models *of* ideal agents," and when we refer to credences and utilities as those agents' beliefs and desires.

Much of our work in formal philosophy involves the manipulation of the model objects (the mathematical structures), by deriving results and interpreting them in terms of the properties of agents. This is what we are doing when we prove a representation theorem, as discussed in section 7.2: we prove a theorem stating that our binary relation $\succeq$ can be represented by a probability measure, P. We extend the model's interpretation to cover this function and its properties: credences are the probabilistic partial beliefs of ideal agents. We can use the rich structure of probability theory to more easily manipulate this model object, and prove all manner of results—about the rationality constraints on partial belief in general, or about particular situations where we fill in the model description with additional details (say, about a decision an agent wants to make).

7.5.4 *Purpose, scope and criticism*

Scientific models are purpose-specific, with restricted domains of applicability. Given what has come before, it is hopefully now plausible to you that our mathematical frameworks in formal philosophy are models too, albeit with normative ingredients. So are they, too, purpose-specific and domain-restricted? In a weak sense of purpose-specificity, this might seem trivial. They are agential models of rationality, built to explore the rationality conditions on partial belief. That fills the basics of "purpose"—it tells us what the model is a model of, and what it is trying to do. But does this purpose also lead to modelling choices that restrict the model's usefulness in answering other questions? Are our models evaluated not on their truthfulness or truthlikeness, but instead on their adequacy for purpose? I think the answer must be yes (because I think we are modelling), but that this has not been sufficiently acknowledged by many philosophers working in FE. So here is one place where it matters that philosophers are (unknowingly) using the tools of modelling without acknowledging their limitations.

From the discussion above, we can discern one difference between the descriptive and normative cases. Normative models have an additional ingredient in the specification of their domain: the audience for normative guidance. They have one ingredient fewer, too: they lack a target of representation. The difference this makes is small, for the purposes of this section. The kinds of inferences we draw from normative and descriptive models are different, and they apply to different objects (in the normative case, the audience; in the descriptive, the target). But in each case, we draw inferences from the model that are intended to apply to some external object. The question is: what constrains this inferential process?

One answer comes from a consideration of purpose. Philosophers working with normative models put them to a number of different purposes. They might build a model to test a candidate norm—seeing what prescriptions emerge from a model employing it, and testing these against intuitions about what counts as rational. Such a modelling purpose will set implicit criteria for success: a good model for this purpose is one which can easily generate results to be tested in these sorts of cases. Or, they might aim to deploy norms for evaluative purposes (rather than testing whether they are in fact norms). Here, success will involve generating clear evaluative criteria. Finally, we may shift from evaluative to prescriptive normativity and seek to provide action-guidance. Success here will look quite different: action-guiding models need to be usable by those they provide guidance for, and this usability criterion may diverge significantly from the prior two.

Consider a Bayesian decision theory model. Agents in this model have probabilistic credences, they update their beliefs by conditioning on new evidence, and they make decisions by maximising subjective expected utility. These models do well on the first two purposes discussed above: they are simple to use to generate decisions in test cases like the Allais and Ellsberg scenarios, and establish clear criteria for rational belief, preference and decision. They are not very helpful for action-guidance, however. The process of eliciting any real person's attitudes and representing them as utilities and probabilities is onerous. As is often noted, Bayesianism demands too much of real agents. But note that this isn't a problem if the purpose is norm-testing or evaluation. It is only a problem if the modeller intends the model to be used for action-guidance.

A summary account of normative models

One of the main tasks of the philosophy of modelling literature has been to account for how it is that models function. An example of such an account is the DEKI account of Frigg and Nguyen (2016), which builds on the notion of representation-as. I will briefly outline that account here and sketch how a parallel account could be developed for normative models. (I use the DEKI account as I am most familiar with it; I see no reason that the same work could not be done for other major accounts of model representation.)

Recall the plastic model of the water molecule, and as above let us understand it as a Z-representation: an object equipped with an interpretation. DEKI stands for Denotation, Exemplification, Keying-up and Imputation, the four elements of how such a model functions.

First, the model *denotes* some target system under study: the H_2O molecule. (This makes it a representation-*of* water.) "Denotation" is used to describe this relation because it is an intentional act of a modeller, and because it contains no more content than the identification of the target. Not all models have targets, and so in some cases the D element is not in use. Second, the model *exemplifies* some relevant properties (technically, I-exemplifies—I'm not responsible the for the acronym). This is a technical notion, that precisifies the notion of "highlighting" a property as relevant that I discussed informally above. To exemplify a property is to have it and refer to it, which accomplishes the "highlighting". This is accomplished by the interpretation, hence the "I" in I-exemplify. Now the basic idea is that we want to investigate the model-system, learn about these properties and then infer something about the target, H_2O. But in many cases, we don't want to simply impute the exemplified properties of the model to the target. Some are going to be artefacts of the modelling process, and others will be subject to distortions introduced by idealisations. The model therefore comes with a *key*, which is the third element of DEKI. A key specifies how exemplified properties represent properties of the target. For example, some properties might be intended to be read as approximate: if the angle between the two arms of the V is 105°, the key might tell us to infer that H_2O has an angle of approximately 105°. Other properties, like the elemental composition of the molecule, may be directly inferred to the target. The final step is to *impute* these (potentially modified) keyed-up properties to the target system: the model tells us that water molecules involve one hydrogen and two oxygen atoms, that they have a certain structure, and so forth.

The DEKI account tells us *in virtue of what* models represent—it provides a semantics for modelling. As representation is taken to be the core of how (many, perhaps most) scientific models function, it is thereby an account of model functioning. My aim in this section is to sketch a parallel account for normative philosophical models. As we have seen, it is not prima facie obvious that representation is the core of normative modelling, so the DEKI account itself will not do.

It can function as a starting point, however, since normative models do have *some* representational function, as I argued above. Normative models are representations-as, object-interpretation pairs. This they have in common with descriptive scientific models. Normative models are not representations-of anything, however. But this too is something in common with some scientific models: the target-less models. On the DEKI account, the study of target-less models focuses on the model-object: its properties, and how these I-exemplify properties of the kind of thing it Z-represents. The focus is on interpretation and exemplification, because there is no target denoted and therefore nothing to impute properties to. I therefore propose to work through the "silent" elements of the DEKI account (the D, K and I) to build out my account of normative models.

We begin with "D": Normative models do not denote anything, as they do not have representational targets. What they do often have is an *audience*, the intended recipients of normative guidance. The relation of the normative model to this audience is similar in one important respect to the denotation relation that holds between targeted descriptive models and their targets: it is fixed by the intentionality of the modeller. Note that "denotation" in the DEKI system is understood as entirely conventional; a matter simply of the intentions of the modeller. *Good* modellers will be able to judge which models are good fits for which targets, but that is a matter of the success conditions of modelling, not of what it is to *be* a model. Similarly, the good normative modeller will be able to judge which models fit well with which normative audiences, and here that judgement will be fixed by the modeller's understanding of the deontic scope of the norms they are working with: the people the modeller believes they are norms *for*. However, this is not, I think, a condition for *being* a normative model, but for being a successful one.

Skipping now to the "I": Normative models do not impute properties to targets (they don't have those), nor to their audiences. Claiming that property is true of the audience isn't the purpose of a normative model. Instead, modellers *generate claims* about duties or evaluative standards applicable to the audience, by analysing the properties of the model. As noted before, these can be generated with the intention of genuinely serving as norms for the audience, or as part of a process of testing the plausibility of putative norms by checking the recommendations against intuition.

This process of drawing inferences will again be skilful, requiring a Key. First, there may be different claim-generating schemas, such as "model-agents have property X, therefore you should act in such a way as to come to have X," and "model-agents have property Y, therefore you will be evaluated against Y as a normative standard." The particular schemas in operation will depend on the purposes of the modeller, and the form of the norms. Second, how the properties of the model are translated into claims may be modulated by the Key. On the one hand, we may wish to export a normative demand to approximate a property that the model has precisely, much in the way that a scientist might take only the sign of a result to be relevant to the target. On the other, normative models will contain non-normative idealisations and so, as in the scientific case, care will be needed to determine which properties of the model are "genuine" and which are artefacts of the modelling process. Rather than "genuine" meaning faithful representation, as in the science case, now it means something like feature of our normative account.

Normative models have a four-step process that parallels DEKI. (1) They have an audience, which is fixed by the intentions of the modeller. (2) The models exemplify certain properties, which are fed to (3) the model's key, which tells us how to (4) generate normative claims for the audience. The purpose of providing such a summary account is the same here as in the descriptive case: it ties together the various aspects of model-functioning, and provides a framework for future work on normative modelling.

7.6 ASIDE: WHAT ELSE COULD WE BE DOING?

Some readers (though perhaps not those who have gotten this far!) might think: "Of course it is modelling, what else could it have been?" Why spill so much ink on the topic?

There are other methodologies that we might be using. Not all of science is modelling, and I don't think all of philosophy is either. Godfrey-Smith (2007) distinguishes the model-based "strategy" of science from an alternative, more direct, method of theorising. He contrasts two projects in late-twentieth century biology: Maynard-Smith and Szathmáry's *The Major Transitions in Evolution* (1995) and Leo Buss's *The Evolution of Individuality* (1987). The former is an exercise in model-based science, introducing many different models to isolate and discuss various causal mechanisms. The latter employs no models at all; instead, Buss examines actual organisms, in their actual circumstances. This work is close to the data, and involves studying real rather than fictional systems. It is synoptic, making progress by systematising knowledge (Godfrey-Smith compares Buss's work to Darwin's).

So, even if naturalistic philosophy recommends using scientific methods, these needn't be modelling. This direct method, however, is not a good description of formal epistemology. We don't work from close attention to real agents, and not merely because we have normative aims. Consider Cassam's (2019) *Vices of the Mind*. This is epistemology done "from the ground up"—the theory of epistemic vices is built from a close examination of real cases involving real people, but the theory is put to normative ends. It is manifestly unlike the formal epistemology discussed here.

One difference, of course, is that Cassam's work is not formal. But not all formal work is modelling either. Keefe (2000) is insistent that her supervaluationist work on vagueness is not intended as a model—an idealised, indirect representation of the linguistic phenomenon. She aims at a true description of the phenomenon of vague language (Keefe, 2000, Ch. 1). Accordingly, her methodology is different—it isn't idealised in the sense I have discussed here—and so are the success conditions for her work. As she notes, it is not open to Keefe to tell us to disregard certain parts of her mathematical framework as artefacts, or to isolate her account of vagueness from other accounts of linguistic functioning. Her work *is* open to refutation by counterexample, by design.

Modelling is a method rather than a goal. It therefore doesn't conflict with traditional philosophical *project* of conceptual analysis, or species thereof like Carnapian explication. As noted above, models are used to isolate mechanisms or concepts for particular study (Wimsatt, 2007, p. 15). Models can therefore support conceptual analysis or Carnapian explication by providing an isolated testing ground for a new concept. Similarly, multiple idealised models can support the development of fuller theories (which we might want not to be simplified or distorted), through the examination of results on which all models agree.

Given this, other readers might say: "if modelling is one strategy among many, then sure modelling requires a conscious intentional stance to one's work. If the people you describe don't think they're modelling, then they aren't." I think that this would hold only if all involved understood what the different methods were, and were reflective about their method as they worked. Neither is always true: I think some philosophers have been confused about their methods, or have lacked the conceptual machinery to realise that what they're doing is modelling. The aim of this chapter is to provide such machinery, in order that we can realise when we are modelling and to facilitate the choice not to model.

7.7 METHODOLOGY AND INFERENCE

This brings us to our discussion of the methodology of normative modelling. As we just saw, criticisms of normative models must take heed of their purposes. They must also pay attention to the model's key. To criticise a Bayesian model for properties that skilled users know to disregard—such as logical omniscience, or instant computation—

is to misunderstand the methodology of modelling. That said, if a particular result depends in an important way on a property keyed as an artefact, it is similarly a methodological error to make use of that property—either imputing it descriptively to a target in the descriptive case, or making a normative inference using it in the normative case.

This reflection allows us to formulate methodological constraints on the kinds of inferences we can draw from normative models. I will present a few important lessons about inferences that we shouldn't draw, along with discussions of cases where they have been drawn, to highlight the implications of coming to regard FE as modelling.

- Property X appears in our best account of rational partial belief. Therefore, agents are rationally required to have property X. (The "argument for probabilism from representation theorems" employs this move—e.g., (Maher, 1993), and see (Hájek, 2008; Konek, 2019) for discussion.)

  One we replace the term "account" with "model" it becomes clear we need to be careful. In the descriptive case, realist inferences from discovered properties of the model to the target must be motivated with reference to their (in)dependence on idealising assumptions. Similarly, in the normative case, not all properties in the model are going to count as normative. Ignorance of a model's key makes it very difficult to cogently criticise. This leads to a methodological norm for modellers: be clear about what you regard as an artefact, and what you intend to be imputed to the target.

- Property X appears in your account. Property X is absurd, so your account is false. (Glymour's (1980) argument against Bayesianism repeatedly deploys this move.)

  As above, we now see that useful models may contain worrisome properties, which must not be imputed to the target. Sometimes we will need to avoid applying the model to cases where that property would do important work.

- Your account doesn't work in case Y. Y is a counterexample, so your account is false. (Very common, but e.g., the argument against imprecise probabilism in (Elga, 2010) has this form.)

  Models have a domain of applicability, so each "counterexample" must be checked against this domain. Objections irrelevant to the model's intended purpose have no bite. Instead, they motivate for a different model to be developed (perhaps to handle just those cases, or to expand the scope). Working out the boundaries of applicability for different philosophical models is a research area deserving of more attention.

As these moves are common, there is an important debate to be had about which bits of formal philosophy are modelling, and which are not. However, "it is just a model" should not be a get-out-of-jail-free card against objections (Keefe (2000, pp. 49-56) accuses some vagueness theorists of using it this way). This reinforces the need for clarity on the purpose and context guiding the modelling, and its key.

7.7.1 *Securing normative inferences*

I now want to return to the question raised at the end of the section (7.5.2) on idealisation. How can we know that our normative inferences are "secure" in the face of their dependence on non-normative idealisations?

The problem concerning us here is that some of our normative conclusions depend necessarily on these assumptions. Consider Probabilism, the claim that one must have partial beliefs that are probabilistically representable. But, as we have seen, probabilistic representability requires that one's partial beliefs are complete and continuous.[15] In the scientific case, results which depend necessarily and sensitively on heuristic idealisations are generally regarded as artefacts and not taken to inform us about the target. Supposing for the moment that completeness and continuity are neither normative standards nor approximately true descriptions of partial belief, what does that mean for Probabilism? More generally, what can we say about when our normative inferences from models are secure?

Norm in, norm out

The easiest case is this: we construct a model in which all the idealisations employed are normatively justified. The agent-representation that results would differ from real agents only in ways that are nor-

[15] Properties of this sort show up in all of the representation theorems for credences that I am aware of. Without completeness, it is impossible to ensure that each proposition can be assigned a unique number in the way that credences do, and without continuity (or one of its any variants) one cannot get the rich structure of the real number line—in particular, the fact that when we have two real numbers, we can always find a third that lies between them.

mative. Then, it would be clear that the results generated by the model can be used to generate normative claims: norm in, norm out. But this is a limiting case that isn't that helpful— a purely normative model would be like a descriptively accurate scientific model. Its inferences would be secure, but it would not in truth be a "model". Models gain their usefulness from their ability to simplify through distortion and abstraction.

Approximate norms

Next, consider a model which employs some number of normative idealisations and also one non-normative idealisation, which is justified by a plausible negligibility argument, perhaps bolstered by a domain restriction to the cases where it works best. (Perhaps we have a group of people who have nearly complete partial beliefs over some algebra of relevant propositions.) Now it seems we are on good ground. Our model is idealised, and we may wish to employ sensitivity analysis to determine whether the precise nature of the idealisation is generating any artefacts. But if the negligibility argument is sound, it provides good reason to think that we can find a model which captures the remaining structural and causal properties of the system without introducing sensitive dependence on the idealised factor.

Now suppose we generate some results from working with the model—can we draw secure normative inferences for the model's audience? I think the answer is yes, but with a suitable understanding of what it is that models can achieve. Scientific models work well under these conditions, but their use involves an acknowledgement of their limitations and fallibility. Models focus on what's most important, but in so doing they sacrifice precision. Their value is in capturing the main features of a system's behaviour, driven by the most important underlying factors. In complex systems this is remarkable, but it comes at the cost of precision. Philosophers often want their norms to be categorical, which this would block. Instead we may need to learn to present our results as "approximately normative," or as candidate norms that need to be confirmed by less idealised methods.

Formal epistemologists do not always do this. Recall that in our partial belief model, we ended up with a long list of assumptions about the credibility relation, if we want to prove a Joyce/Villegas-style representation theorem: it must be monotonic, separable, transitive, complete and continuous. The chief normative result from this model that I considered was Probabilism, the thesis that if one's beliefs cannot be probabilistically represented then one is irrational. Few authors in FE have taken the time to carefully determine whether this normative conclusion is purely a function of normative assumptions, or to examine the potential complications of deploying non-normative assumptions in its derivation. This, I think, is a mistake.

The first three of the assumptions about credilibility, I introduced as norms; the latter two are what decision theorists call "structural assumptions". That is to say that they are neither norms nor descriptive claims about real agents' partial beliefs, but instead are assumptions about the mathematical structure of the model object, introduced to make the analysis easier.[16] In our language, we can think of these as idealisations justified as domain-restrictions, or heuristically.

For example, economists are famously cavalier about completeness of preference, regarding it as delineating the scope of the problem they are concerned with.[17] (This is implausible, and much criticised (e.g., by Joyce, 1998, pp. 98-103), but for the moment let us take the point to be: there is a practice of justifying structural assumptions in language that is recognisable to the philosophy of scientific modelling.) A parallel justification for the case of credibility would be to assert that we are only modelling cases where an agent considers a limited number of propositions and does, as a matter of fact, make comparative judgements of likelihood about all pairs of propositions. This can succeed in making the idealisation innocuous, at the cost of reducing the model's scope (perhaps drastically).

A less restrictive justification would be to regard it as heuristic: a simplification for the time being. Something like this thought is present in the decision theory/FE literature in the form of the "co-

[16] The term "structural assumption" appears to come from the measurement theory literature. Krantz et al. (1971) say: "Nonnecessary axioms are frequently referred to as *structural* because they limit the set of structures satisfying the axiom system to something less than the set determined by the representation theorem." (They are here referring to a more primitive representation theorem in which one establishes that the basic thing being measured—i.e. the attitude of partial belief—can be represented with an ordinal structure.) They are using "structure" and "satisfy" in a set-theoretic and model-theoretic sense. The key point is that this reduction is meant to select a set of structures which are easy to work with. So in our case we reduce the set from those that are merely monotonic, separable, and transitive to the special subset which are also continuous and complete and thereby representable by a probability function.

[17] See for example (Arrow, 1966, p. 225), (Luce and Raiffa, 1957, p. 287), as well as a more modern example in (Gilboa, 2009, pp. 51-2). Gilboa describes completeness as normative, in the sense of being an injunction to the decision-maker: face-up to your decisions! But in the context of his presentation of preference as derived from choice, and of some choice in fact being made from a set of options, I think it is natural to describe his move as a domain-restriction.

herent extendibility" thesis. This thesis states that there is nothing irrational about not being able to make a judgement about which of two propositions is more likely, but there is an important constraint contained in completeness nonetheless. A rational agent's credibility ranking must be such that it is possible to extend the relation to one that is complete, without violating any of the core rationality axioms (transitivity, monotonicity, separability) in the process (see Jeffrey, 1992, p. 85 and Joyce, 1998, p. 103 for discussion). This amounts to a complex key for interpreting the normative results of the model.

Complex dependency

More commonly, normative results (like Probabilism, and Savage's decision theory) depend on a complex mixture of normative and non-normative assumptions. In this general case, I take the methodological import of this chapter to be: with great power (idealisation) comes great responsibility (sensitivity and robustness analysis, humility in the face of model pluralism, careful attention to the model's purpose and domain). Modelling is a difficult business, and philosophers have thus far rarely exhibited the careful analysis required to extract secure inferences from their models.[18]

[18] Some parts of the wider formal epistemology community already act in roughly the manner that I recommend here. In formal social epistemology, where much work consists of model building and application, there is a conscious effort to attend to good modelling methodology, in a way that aligns with my recommendations. For example see the discussion of (the lack of) stability analysis in Zollman (2010) by Rosenstock, O'Connor, and Brunner (2017) and Frey and Šešelja (2018). This work, however, is largely descriptive rather than normative.

7.8 CONCLUSION

Once we accept that what we are doing is modelling, the implications for our philosophical practice are wide-reaching. One immediate impact is that it shows a certain fruitlessness to the current debate between precise probabilists (PP) and imprecise probabilists (IP). What is at stake in that debate is a norm (the permissibility of ambiguity aversion). But much of the debate takes place at the level of model results, which are complex functions of normative and non-normative idealisations. If these models employ different idealisations and were built for different purposes, we now see that they may not be easy to compare.

Consider the completeness property again. One route to IP is via coherent extendibility: you accept that credibility is not complete, but require that it be coherently extendible. The class of all coherent completions of a credibility relation generates a set of probability functions, which are then taken to be the imprecise representor of the agent's partial beliefs. For one purpose, the precision of a PP model might be favoured and regarding completeness as a negligibility or domain idealisation may introduce no great difficulties. In another case, where completeness would obstruct her purpose, the modeller might use an IP model instead. This is a natural state of affairs once we see that we are modelling. Indeed, it was always permissible for the Impreciser because of a further confusing fact about this debate: while PP and IP disagree over the norms of rationality, the set of IP models includes the set of PP models, and so Imprecisers are free to claim that certain contexts and purposes support the use of a PP model. Precisers, on the other hand, are committed to the claim that only precise models are acceptable. In the debate, their strategy must therefore be to block the permissibility of ever using an IP model. In order to do so, they must either conclusively reject the norm at stake; or they must identify goals/purposes so universal that no model that does not accommodate them can succeed, and then show that no IP model can do so.

The precision debate itself is borne of a sense that there must be a single, true normative account of partial belief. This is an admirable goal, but we must not mistake these two models, idealised and distorted as they are, for candidates for such an account. One important lesson from scientific modelling is that a multiplicity of models is no bad thing! Each is likely to do best on its "home turf", and each will have different lessons for us about partial belief. Careful study of the characteristic problems for each model will help us to identify their home turfs and the boundaries of their domains of applicability. With these in hand we can turn to more important issues than fighting about whether Precise or Imprecise Probabilism is correct, such as looking at areas where neither does well. Here, a new model is needed.

Does this model pluralism commit us to antirealism? In truth, I am not sure. This may be a concern for philosophers more used to "direct" theorising about their domains. While I am not personally concerned if my conclusion is antirealist, there may be a careful path toward realism—it is certainly not the case that every modeller and philosopher of modelling is an antirealist. What they are, however, is very careful about elevating claims about the content of models to claims about the content of reality. Careful attention to "robust results" can reveal what is common between disagreeing models, which in turn may be candidates for realist inference. Careful de-idealisation is another route, though one which may prove intractable. In either case, formal epistemologists will benefit from attending to the philosophy of scientific modelling.

BIBLIOGRAPHY

Armstrong, J. Scott (2001). "Combining Forecasts". In: *Principles of Forecasting*. Springer, Boston, MA, pp. 417–439. DOI: 10.1007/978-0-306-47630-3_19.

Arrow, Kenneth J. (1966). "Exposition of the Theory of Choice under Uncertainty". In: *Synthese* 16, pp. 253–69.

Aspinall, Willy (2010). "A Route to More Tractable Expert Advice". In: *Nature* 463.21, pp. 294–25.

Aumann, Robert J. (1976). "Agreeing to Disagree". In: *The Annals of Statistics* 4.6, pp. 1236–1239.

Australian Bureau of Meteorology (2017). *Forecast Verification*. https: //web.archive.org/web/20171125111801/ https://www.cawcr.gov.au/projects/verification/.

Babic, Boris (2019). "A Theory of Epistemic Risk". In: *Philosophy of Science* 86.3, p. 550. DOI: 10.1086/703552.

Baccelli, Jean and Rush T. Stewart (ms). "Support for Geometric Pooling". In: *manuscript*.

Bender, Morris A., Thomas R. Knutson, Robert E. Tuleya, Joseph J. Sirutis, Gabriel A. Vecchi, Stephen T. Garner, and Isaac M. Held (2010). "Modeled Impact of Anthropogenic Warming on the Frequency of Intense Atlantic Hurricanes". en. In: *Science* 327.5964, pp. 454–458. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1180568.

Betz, Gregor (2009). "Underdetermination, Model-Ensembles and Surprises – On the Epistemology of Scenario-Analysis in Climatology". In: *Journal for General Philosophy of Science* 40, pp. 3–21.

— (2013). "In Defence of the Value Free Ideal". In: *European Journal for Philosophy of Science* 3, pp. 207–20.

Blake, E.S., E.N Rappaport, J.D. Jarrell, and C.W. Landsea (2005). "The Deadliest, Costliest, and Most Intense United States Tropical Cyclones". In:

Booth, Ben B. B., Nick J. Dunstone, Paul R. Halloran, Timothy Andrews, and Nicolas Bellouin (2012). "Aerosols Implicated as a Prime Driver of Twentieth-Century North Atlantic Climate Variability". en. In: *Nature* 484.7393, pp. 228–232. ISSN: 1476-4687. DOI: 10.1038/nature10946.

Bovens, Luc and Stephan Hartmann (2003). *Bayesian Epistemology*. Oxford: Oxford University Press.

Bradley, Richard (2005). “Radical Probabilism and Bayesian Conditioning”. en. In: *Philosophy of Science* 72.2, pp. 342–364. ISSN: 0031-8248, 1539-767X. DOI: 10.1086/432427.

— (2007). “Reaching a Consensus”. In: *Social Choice and Welfare* 29.4, pp. 609–632. ISSN: 0162-1459. DOI: 10.1007/s00355-007-0247-y.

— (2017). *Decision Theory with a Human Face*. Cambridge University Press. ISBN: 978-1-107-00321-7.

— (2018). “Learning from Others: Conditioning versus Averaging”. In: *Theory and Decision* 85, pp. 5–20.

— (ms). *Decision-Making with Catastrophe Models*.

Bradley, Richard, Franz Dietrich, and Christian List (2014). “Aggregating Causal Judgments”. In: *Philosophy of Science* 81.4, pp. 491–515. DOI: 10.1086/678044.

Bradley, Richard, Casey Helgeson, and Brian Hill (2017). “Climate Change Assessments : Confidence, Probability, and Decision”. In: *Philosophy of Science* 84.3, pp. 500–522. DOI: 10.1086/692145.

Brier, Glenn W (1950). “Verification of Forecasts Expressed in Terms of Probability”. In: *Monthly Weather Review* 78.1, pp. 1–3.

Buchak, Lara (2013). *Risk and Rationality*. Oxford University Press. ISBN: 978-0-19-175904-8.

Cartwright, Nancy (1989). *Nature's Capacities and Their Measurement*. Oxford University Press.

Cassam, Quassim (2019). *Vices of the Mind, From the Intellectual to the Political*. Oxford University Press.

Christensen, D. (2007). “Epistemology of Disagreement: The Good News”. en. In: *Philosophical Review* 116.2, pp. 187–217. ISSN: 0031-8108, 1558-1470. DOI: 10.1215/00318108-2006-035.

Churchman, C. West (1948). “Statistics, Pragmatics, Induction”. In: *Philosophy of Science* 15.3, pp. 249–268. ISSN: 0031-8248. DOI: 10.1086/286991.

Clemen, Robert T. (1989). “Combining Forecasts: A Review and Annotated Bibliography”. In: *International Journal of Forecasting* 5.4, pp. 559–583. ISSN: 0169-2070. DOI: 10.1016/0169-2070(89)90012-5.

Colyvan, Mark (2013). “Idealisations in Normative Models”. en. In: *Synthese* 190.8.

Cooke, Roger (1999). *Experts in Uncertainty*. Oxford: Oxford University Press.

— (2018). *Structured Expert Judgement Applications*.

Cooke, Roger and L. H. J Goossens (2000). “Procedures Guide for Structured Expert Judgement in Accident Consequence Modelling”. In: *Radiation Projection Dosimetry* 90.3, pp. 303–309.

Council, US National Research. *About Our Expert Consensus Reports*. http://dels.nas.edu/global/consensus-report.

DeGroot, Morris H. (1988). "A Bayesian View of Assessing Uncertainty and Comparing Expert Opinion". In: *Journal of Statistical Planning and Inference* 20, pp. 295–306.

Diaconis, Persi and Sandy L. Zabell (1982). "Updating Subjective Probability". In: *Journal of the American Statistical Association* 77.380, pp. 822–30.

Dietrich, Franz (2010a). "Bayesian Group Belief". In: *Social Choice and Welfare* 35.4, pp. 595–626. ISSN: 01761714. DOI: 10.1007/s00355-010-0453-x.

— (2010b). "Bayesian Group Belief". In: *Social Choice and Welfare* 35.4, pp. 595–626. DOI: 10.1007/s00355-010-0453-x.

Dietrich, Franz and Christian List (2016a). "Probabilistic Opinion Pooling". In: *Oxford Handbook of Probability and Philosophy*. Ed. by Alan Hajek and Christopher Hitchcock. Oxford: Oxford University Press. DOI: 10.1093/oxfordhb/9780199607617.013.37.

— (2016b). "Probabilistic Opinion Pooling". In: *Oxford Handbook of Probability and Philosophy*. Ed. by Alan Hajek and Christopher Hitchcock. Oxford: Oxford University Press. DOI: 10.1093/oxfordhb/9780199607617.013.37.

— (2017). "Probabilistic Opinion Pooling Generalized. Part One: General Agendas". In: *Social Choice and Welfare* 48.4, pp. 747–786.

Dietrich, Franz, Christian List, and Richard Bradley (2016). "Belief Revision Generalized: A Joint Characterization of Bayes' and Jeffrey's Rules". In: *Journal of Economic Theory* 162, pp. 352–371. ISSN: 0022-0531. DOI: 10.1016/j.jet.2015.11.006.

Dietrich, Franz and Kai Spiekermann (2013). "Epistemic Democracy with Defensible Premises". In: *Economics and Philosophy* 29.1, pp. 87–120.

Douglas, Heather (2009). *Science, Policy and the Value-Free Ideal*. Pittsburgh: University of Pittsburgh Press.

Douven, Igor (2018). "Scoring in Context". In: *Synthese*, pp. 1–16. DOI: https://doi.org/10.1007/s11229-018-1867-8.

Easwaran, Kenny, Luke Fenton-Glynn, Christopher Hitchcock, and Joel D. Velasco (2016). "Updating on the Credences of Others: Disagreement, Agreement, and Synergy". en. In: *Philosopher's Imprint* 16.11.

Elga, Adam (2007). "Reflection and Disagreement". en. In: *Nous* 41.3, pp. 478–502.

— (2010). "Subjective Probabilities Should Be Sharp". In: *Philosopher's Imprint* 10.5, pp. 1–11.

Elgin, Catherine Z. (1983). *With Reference to Reference*. Indianapolis and Cambridge: Hackett.

Elgin, Catherine Z. (2009). “Exemplification, Idealization, and Understanding”. In: *Fictions in Science: Essays on Idealization and Modeling*. Ed. by Mauricio Suarez. London: Routledge, pp. 77–90.

— (2010). “Telling Instances”. In: *Beyond Mimesis and Nominalism: Representation in Art and Science*. Ed. by Roman Frigg and Matthew C. Hunter. New York: Springer, pp. 1–18.

Ellsberg, Daniel (1961). “Risk, Ambiguity, and the Savage Axioms”. In: *Quarterly Journal of Economics* 75.4, pp. 643–669.

Emanuel, Kerry A. (2005). *Divide Wind - The History and Science of Hurricanes*. New York: Oxford University Press.

Enfield, David B. and Luis Cid-Serrano (2006). “Projecting the Risk of Future Climate Shifts”. In: *International Journal of Climatology* 26.7, pp. 885–895. ISSN: 0899-8418. DOI: 10.1002/joc.1293.

Eva, Benjamin and Stephan Hartmann (2019). “On the Origins of Old Evidence”. In: *Australasian Journal of Philosophy* forthcoming in print, pp. 1–14.

Eva, Benjamin, Stephan Hartmann, and Soroush Rafiee Rad (2019). “Learning from Conditionals”. en. In: *Mind*, pp. 1–48. ISSN: 0026-4423, 1460-2113. DOI: 10.1093/mind/fzz025.

FCHLPM (2007). *Report to the Florida House of Representatives Comparison of Hurricane Loss Projection Models*.

— (2015). *Current Year 2015 Modeler Submissions*. https://www.sbafla.com/method/ModelerSubmissions/CurrentYear2015ModelerSu

French, Simon (1980). “Updating of Belief in the Light of Someone Else’s Opinion”. en. In: *Journal of the Royal Statistical Society. Series A (General)* 143.1, p. 43. ISSN: 00359238. DOI: 10.2307/2981768.

Frey, Daniel and Dunja Šešelja (2018). “What Is the Epistemic Function of Highly Idealized Agent-Based Models of Scientific Inquiry?” en. In: *Philosophy of the Social Sciences* 48.4, pp. 407–433. ISSN: 0048-3931. DOI: 10.1177/0048393118767085.

Frigg, Roman and Stephan Hartmann (2018). “Models in Science”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2018. Metaphysics Research Lab, Stanford University.

Frigg, Roman and James Nguyen (2016). “The Fiction View of Models Reloaded”. In: *Monist* 99.3, pp. 225–242. DOI: 10.1093/monist/onw002.

— (2017). “Models and Representation”. en. In: *Springer Handbook of Model-Based Science*. Ed. by Lorenzo Magnani and Tommaso Bertolotti. Dordrecht, Heidelberg, London and New York: Springer, pp. 49–102.

Frigg, Roman, Erica Thompson, and Charlotte Werndl (2015). “Philosophy of Climate Science Part I: Observing Climate Change”. In: *Philosophy Compass* 10.12, pp. 953–964. ISSN: 1747-9991. DOI: 10.1111/phc3.12294.

Gaifman, Haim (1988). “A Theory of Higher Order Probabilities”. In: *Causation, Chance, and Credence*. Vol. 1. Kluwer, pp. 191–220.

Garber, Daniel (1984). “Old Evidence and Logical Omniscience in Bayesian Confirmation Theory”. In: *Testing Scientific Theories*. Ed. by John Earman. University of Minnesota Press.

Gärdenfors, Peter and Nils-Eric Sahlin (1982). “Unreliable Probabilities, Risk Taking and Decision Making”. In: *Synthese* 53, pp. 361–386.

Genest, Christian (1984a). “A Characterization Theorem for Externally Bayesian Groups”. In: *The Annals of Statistics* 12.3, pp. 1100–1105.

— (1984b). “A Conflict between Two Axioms for Combining Subjective Distributions”. In: *Journal of the Royal Statistical Society, Series B* 46, pp. 403–405.

Genest, Christian and Mark J. Schervish (1985). “Modeling Expert Judgments for Bayesian Updating”. en. In: *The Annals of Statistics* 13.3, pp. 1198–1212. ISSN: 0090-5364. DOI: 10.1214/aos/1176349664.

Genest, Christian and James V. Zidek (1986). “Combining Probability Distributions: A Critique and an Annotated Bibliography”. In: *Statistical Science* 1.1, pp. 114–135. ISSN: 08834237. DOI: 10.1214/ss/1177013825.

Giere, Ronald N (1988). *Explaining Science*. Science and Its Conceptual Foundations. Chicago: University of Chicago Press.

— (2004). “How Models Are Used to Represent Reality”. In: *Philosophy of Science* 71, pp. 742–52.

Gilboa, Itzhak (2009). *Theory of Decision under Uncertainty*. Cambridge: Cambridge University Press.

Gilboa, Itzhak and Massimo Marinacci (2013). “Ambiguity and the Bayesian Paradigm”. In: *Advances in Economics and Econometrics: Theory and Applications*. Ed. by D Acemoglu, M Arellano, and E Dekel. Cambridge: Cambridge University Press.

Gilboa, Itzhak and David Schmeidler (1989). “Maxmin Expected Utility with Non-Unique Prior”. In: *Journal of Mathematical Economics* 18.2, pp. 141–153. ISSN: 0415324947. DOI: 10.1016/0304-4068(89)90018-9.

Glymour, Clark N. (1980). *Theory and Evidence*. Princeton: Princeton University Press.

Godfrey-Smith, Peter (2006). “Theories and Models in Metaphysics”. en. In: *The Harvard Review of Philosophy* 14.1, pp. 4–19. ISSN: 1062-6239. DOI: 10.5840/harvardreview20061411.

— (2007). “The Strategy of Model-Based Science”. en. In: *Biology & Philosophy* 21.5, pp. 725–740. ISSN: 0169-3867, 1572-8404. DOI: 10.1007/s10539-006-9054-6.

Godfrey-Smith, Peter (2012). "Metaphysics and the Philosophical Imagination". en. In: *Philosophical Studies* 160.1, pp. 97–113.

Goldman, Alvin I (2001). "Experts: Which Ones Should You Trust?" In: *Philosophy and Phenomenological Research* 63.1, pp. 85–110. ISSN: 0031-8205. DOI: 10.1111/j.1933-1592.2001.tb00093.x.

Good, I.J. (1950). *Probability and the Weighing of Evidence*. London: Charles Griffin.

— (1952). "Rational Decision". In: *Journal of the Royal Statistical Society, Series B* 14, pp. 107–114.

Goodman, Nelson (1976). *Languages of Art*. Indianapolis and Cambridge: Hackett.

Guin, Jayanta (2010). *Understanding Uncertainty*. en. http://www.air-worldwide.com/Publications/AIR-Currents/2010/Understanding-Uncertainty/.

Hájek, Alan (2008). "Arguments For, or Against, Probabilism?" en. In: *The British Journal for the Philosophy of Science* 59.4, pp. 793–819.

Hall, Timothy M. and Stephen Jewson (2007). "Statistical Modelling of North Atlantic Tropical Cyclone Tracks". en. In: *Tellus A: Dynamic Meteorology and Oceanography* 59.4, pp. 486–498. ISSN: 1600-0870. DOI: 10.1111/j.1600-0870.2007.00240.x.

Hancox-Li, Leif (2017). "Idealization and Abstraction in Models of Injustice". In: *Hypatia* 32.2, pp. 329–46.

Hansen, Lars Peter and Thomas J Sargent (1982). *Robustness*. Princeton: Princeton University Press. ISBN: 978-1-4008-2938-5.

Hardwig, John (1985). "Epistemic Dependence". In: *Journal of Philosophy, Inc* 82.7, pp. 335–349. ISSN: 0022-362X.

— (1991). "The Role of Trust in Knowledge". In: *Journal of Philosophy* 88.12, pp. 693–708.

Harsanyi, J (1967). "Games of Incomplete Information Played by 'Bayesian' Players. Part I - The Basic Model". In: *Management Science* 14, pp. 159–182.

Harsanyi, John C. (1968a). "Games of Incomplete Information Played by 'Bayesian' Players. Part II - Bayesian Equilibrium Points". In: *Management Science* 14.5, pp. 320–334.

— (1968b). "Games of Incomplete Information Played by 'Bayesian' Players. Part III - The Basic Probability Distribution of the Game". In: *Management Science* 14.7, pp. 486–502.

Heal, G and A Milner (2014). "Uncertainty and Decision Making in Climate Change Economics". In: *Review of Environmental Economics and Policy* 8, pp. 120–137.

Helgeson, Casey, Richard Bradley, and Brian Hill (2018). "Combining Probability with Qualitative Degree-of-Certainty Assessment". In: *Climatic Change* 149.3-4, pp. 517–25.

Hill, Brian (2013). "Confidence and Decision". In: *Games and Economic Behavior* 82, pp. 675–692. ISSN: 0899-8256. DOI: 10.1016/j.geb.2013.09.009.

— (2016). "Incomplete Preferences and Confidence". en. In: *Journal of Mathematical Economics* 65, pp. 83–103. ISSN: 03044068. DOI: 10.1016/j.jmateco.2016.05.007.

— (2019). "Confidence in Beliefs and Rational Decision Making". In: *Economics and Philosophy* 35.2, pp. 223–58. DOI: https://doi.org/10.1017/S0266267118000214.

Hogg, R.V. and E.A. Tanis (2001). *Probability and Statistical Inference*. Sixth. London: Prentice Hall.

Howson, Coling and Peter Urbach (2006). *Scientific Reasoning: The Bayesian Approach*. 3rd. Open Court.

Hughes, R. I. G. (1997). "Models and Representation". In: *Philosophy of Science* 64, S325–S336. ISSN: 0031-8248.

Hurwicz, Leonid (1951). "The Generalised Bayes Minimax Principle: A Criterion for Decision Making Under Uncertainty". In: *Cowles Commission Discussion Paper* 335.

IPCC (2013). *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley. Cambridge: Cambridge University Press.

InsuranceERM (2018). *RMS Responds to AIR's Attack on Hurricane Risk Modelling*. en. https://www.insuranceerm.com/news-comment/rms-responds-to-airs-attack-on-hurricane-risk-modelling.html.

Jeffrey, Richard C (2004). *Subjective Probability: The Real Thing*. Cambridge: Cambridge University Press.

Jeffrey, Richard (1983). *The Logic of Decision*. Second. University of Chicago Press.

— (1992). *Probability and the Art of Judgment*. Cambridge: Cambridge University Press.

Jeffrey, Richard and Michael Hendrickson (1989). "Probabilizing Pathology". en. In: *Proceedings of the Aristotelian Society* 89.1, pp. 211–226. ISSN: 0066-7374, 1467-9264. DOI: 10.1093/aristotelian/89.1.211.

Jehle, David and Branden Fitelson (2009). "What Is the "Equal Weight View"?" en. In: *Episteme* 6.03, pp. 280–293. ISSN: 1742-3600, 1750-0117. DOI: 10.3366/E1742360009000719.

Jewson, Stephen, Enrica Bellone, Thomas Laepple, Kechi Nzerem, Khare Shree, Manuel Lonfat, Adam O Shay, Jeremy Penzer, and Katie Coughlin (2007). "5 Year Prediction of the Number of Hurricanes Which Make U.S. Landfall". en. In: *Hurricanes and Climate Change*. Ed. by J Elsner, p. 37.

Joyce, James M. (1998). "A Nonpragmatic Vindication of Probabilism". In: *Philosophy of Science* 65, pp. 575–603.

Joyce, James M (1999). *The Foundations of Causal Decision Theory*. Online (20. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511498497.

— (2005). "How Probabilties Reflect Evidence". In: *Philosophical Perspectives* 19, pp. 153–178.

Joyce, James M. (2007). "Epistemic Deference: The Case of Chance". en. In: *Proceedings of the Aristotelian Society* 107, pp. 187–206. ISSN: 0066-7373. DOI: 10.1111/j.1467-9264.2007.00218.x.

— (2010). "A Defense of Imprecise Credences in Inference and Decision Making". en. In: *Philosophical Perspectives* 24.1, pp. 281–323. ISSN: 15208583. DOI: 10.1111/j.1520-8583.2010.00194.x.

Kaniovski, Serguei and Alexander Zaigraev (2011). "Optimal Jury Design for Homogeneous Juries with Correlated Votes". en. In: *Theory and Decision* 71.4, pp. 439–459. ISSN: 1573-7187. DOI: 10.1007/s11238-009-9170-2.

Karni, Edi and Marie-Louise Vierø (2013). ""Reverse Bayesianism": A Choice-Based Theory of Growing Awareness". en. In: *American Economic Review* 103.7, pp. 2790–2810. ISSN: 0002-8282. DOI: 10.1257/aer.103.7.2790.

Keefe, Rosanna (2000). *Theories of Vagueness*. Cambridge, New York: Cambridge University Press.

Keeney, Ralph L. and Howard Raiffa (1976). *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. New York: John Wiley & Sons.

Kelly, Thomas (2010). "Peer Disagreement and Higher-Order Evidence". en. In: *Disagreement*. Ed. by Richard Feldman and Ted A. Warfield. Oxford University Press, pp. 111–174. ISBN: 978-0-19-922607-8. DOI: 10.1093/acprof:oso/9780199226078.003.0007.

Kennel, Charles F. (2015). "Global Knowledge Action Network". In: *Proceedings of the Joint Workshop on Sustainable Humanity, Sustainable Nature: Our Responsibility*. Ed. by Partha S. Dasgupta, Veerabhadran Ramanathan, and Marcel Sanchez Sorondo. Vatican City: Pontifical Academy of Sciences, pp. 347–69.

Keynes, John Maynard (1921). *A Treatise on Probability*. London: Macmillan.

Klibanoff, Peter, Massimo Marinacci, and Sujoy Mukerji (2005). "A Smooth Model of Decision Making under Ambiguity". en. In: *Econometrica* 73.6, pp. 1849–1892. ISSN: 1468-0262. DOI: 10.1111/j.1468-0262.2005.00640.x.

Knutson, Thomas et al. (2019). "Tropical Cyclones and Climate Change Assessment: Part I: Detection and Attribution". In: *Bulletin of the American Meteorological Society* 100.10, pp. 1987–2007. ISSN: 0003-0007. DOI: 10.1175/BAMS-D-18-0189.1.

Knutti, Reto (2010). "The End of Model Democracy?" In: *Climate Change* 102, pp. 395–404.

Konek, Jason (2019). "Comparative Probabilities". In: *The Open Handbook of Formal Epistemology*. Ed. by Richard Pettigrew and Jonathan Weisberg. PhilPapers Foundation, pp. 267–348.

Krantz, David H., R. Duncan Luce, Patrick Suppes, and Amos Tversky (1971). *Additive and Polynomial Representations*. Vol. 1. Foundations of Measurement. Academic Press.

Leitgeb, Hannes (2013). "Scientific Philosophy, Mathematical Philosophy, and All That". In: *Metaphilosophy* 44.3, pp. 267–75.

Leitgeb, Hannes and Richard Pettigrew (2010a). "An Objective Justification of Bayesianism I: Measuring Inaccuracy". In: *Philosophy of Science* 77, pp. 201–235.

— (2010b). "An Objective Justification of Bayesianism II: The Consequences of Minimizing Inaccuracy*". en. In: *Philosophy of Science* 77.2, pp. 236–272. ISSN: 0031-8248, 1539-767X. DOI: 10.1086/651318.

Levi, Isaac (1980). *The Enterprise of Knowledge: An Essay on Knowledge, Credal Probability, and Chance*. MIT Press.

Levins, Richard (1966). "The Strategy of Model Building in Population Biology". In: *American Scientist* 54.4, pp. 421–431. ISSN: 0003-0996.

Levinstein, Benjamin Anders (2012). "Leitgeb and Pettigrew on Accuracy and Updating". en. In: *Philosophy of Science* 79.3, pp. 413–424. ISSN: 0031-8248, 1539-767X. DOI: 10.1086/666064.

Lindley, D.V. (1982). "The Improvement of Probability Judgements". In: *Journal of the Royal Statistical Society, Series A* 145.1.

Lloyd, Elisabeth A. (1984). "A Semantic Approach to the Structure of Population Genetics". In: *Philosophy of Science* 51.2, pp. 242–264. ISSN: 0031-8248. DOI: 10.1086/289179.

— (2015). "Model Robustness as a Confirmatory Virtue: The Case of Climate Science". en. In: *Studies in History and Philosophy of Science Part A* 49, pp. 58–68. ISSN: 00393681. DOI: 10.1016/j.shpsa.2014.12.002.

Longino, Helen E. (1990). *Science as Social Knowledge*. Princeton: Princeton University Press.

Luce, R. Duncan and Howard Raiffa (1957). *Games and Decisions: Introduction and Critical Survey*. New York: Wiley.

Mach, Katharine J and Christopher B Field (2017). "Toward the next Generation of Assessment". In: *Annual Review of Environment and Resources* 42, pp. 569–97.

Maher, Patrick T. (1993). *Betting on Theories*. Cambridge University Press.

Mäki, Uskali (2009). "MISSing the World. Models as Isolations and Credible Surrogate Systems". In: *Erkenntnis* 70, pp. 29–43.

Mann, Michael E. and Kerry A. Emanuel (2006). "Atlantic Hurricane Trends Linked to Climate Change". en. In: *Eos, Transactions American Geophysical Union* 87.24, pp. 233–241. ISSN: 2324-9250. DOI: 10.1029/2006EO240001.

Marinacci, M (2015). "Model Uncertainty". In: *Journal of the European Economic Association* 13, pp. 1022–1100.

McConway, K.J. (1981). "Marginalization and Linear Opinion Pools". In: *Journal of the American Statistical Association* 76.374, pp. 410–414.

Met Office (2017). *What Is an Ensemble Forecast? - Met Office*.

Miller, Boaz (2013). "When Is Consensus Knowledge Based? Distinguishing Shared Knowledge from Mere Agreement". In: *Synthese* 190.7, pp. 1293–1316. ISSN: 00397857. DOI: 10.1007/s11229-012-0225-5.

Mills, Charles W (2005). ""Ideal Theory" as Ideology". en. In: *Hypatia* 20.3, pp. 165–185.

Morgan, M Granger (2014). "Use (and Abuse) of Expert Elicitation in Support of Decision Making for Public Policy". In: *PNAS* 111.20, pp. 7176–7184. DOI: 10.1073/pnas.1319946111.

Moss, Sarah (2011). "Scoring Rules and Epistemic Compromise". In: *Mind* 120.480, pp. 1053–1069.

Muir-Wood, Robert and Patricia Grossi (2008). "The Catastrophe Modeling Response to Hurricane Katrina". en. In: *Climate Extremes and Society*. Ed. by Henry F. Diaz and Richard J. Murnane. Cambridge University Press, pp. 296–319. ISBN: 978-1-139-47221-0.

Murphy, Allan H. and Edward S. Epstein (1967). "Verification of Probabilistic Predictions: A Brief Review". In: *Journal of Applied Meteorology* 6.5, pp. 748–755. DOI: 10.1175/1520-0450(1967)006<0748:VOPPAB>2.0.CO;2.

Murphy, J.m, B.b.b Booth, M Collins, G.r Harris, D.m.h Sexton, and M.j Webb (2007). "A Methodology for Probabilistic Predictions of Regional Climate Change from Perturbed Physics Ensembles". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 365.1857, pp. 1993–2028. DOI: 10.1098/rsta.2007.2077.

Musgrave, Alan (1981). "'Unreal Assumptions' in Economic Theory: The F-Twist Untwisted". In: *Kyklos* 34.3, pp. 377–87.

Nau, Robert (2002). "The Aggregation of Imprecise Probabilities". In: *Journal of Statistical Planning and Inference* 105.1, pp. 265–282.

O'Neill, Onora (1987). "Abstraction, Idealization and Ideology in Ethics". In: *Moral Philosophy and Contemporary Problems*. Ed. by J.D.G. Evans. Cambridge: Cambridge University Press.

Oppenheimer, Michael, Brian C O Neill, Mort Webster, and Shardul Agrawala (2007). "The Limits of Consensus". In: *Science* 317.Sept 2007, pp. 1505–1506.

Oppenheimer, Michael, Naomi Oreskes, Dale Jamieson, Keynyn Brysse, Jessica O'Reilly, Matthew Shindell, and Milena Wazeck (2019). *Discerning Experts - The Practices of Scientific Assessment for Environmental Policy*. Chicago: University of Chicago Press.

Oreskes, Naomi and Eric M. Conway (2010). *Merchants of Doubt*. New York: Bloomsbury Press.

Oreskes, Naomi, Kristin Shrader-frechette, and Kenneth Belitz (1994). "Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences". In: *Science* 263.5147, pp. 641–646.

Parker, Wendy S. (2006). "Understanding Pluralism in Climate Modeling". en. In: *Foundations of Science* 11.4, pp. 349–368. ISSN: 1572-8471. DOI: 10.1007/s10699-005-3196-x.

— (2009a). "Confirmation and Adequacy-for-Purpose in Climate Modelling". en. In: *Proceedings of the Aristotelian Society, Supplementary Volumes* 8.

— (2009b). "Does Matter Really Matter? Computer Simulations, Experiments and Materiality". In: *Synthese* 169.3, pp. 483–496.

— (2010). "Whose Probabilities? Predicting Climate Change with Ensembles of Models". en. In: *Philosophy of Science* 77.5, pp. 985–997. ISSN: 0031-8248, 1539-767X. DOI: 10.1086/656815.

— (2011). "When Climate Models Agree: The Significance of Robust Model Predictions". In: *Philosophy of Science* 78.4, pp. 579–600. ISSN: 00318248. DOI: 10.1086/661566.

— (2018). "The Significance of Robust Climate Projections". In: *Climate Modelling - Philosophical and Conceptual Issues*. Ed. by Elisabeth A. Lloyd and Eric Winsberg. Palgrave Macmillan. ISBN: 978-3-319-65057-9.

Paul, L.A. (2012). "Metaphysics as Modeling: The Handmaiden's Tale". en. In: *Philosophical Studies* 160.1, pp. 1–29.

Peirce, Charles S. (1878). "The Probability of Induction". In: *The Popular Science Monthly* XII.

Petersen, Arthur (2012). *Simulating Nature*. Second. Florida: CRC Press.

Pettigrew, Richard (2016). *Accuracy and the Laws of Credence*. Oxford: Oxford University Press.

Pokempner, SJ and EL Bailey (1970). “Sales Forecasting Practices: An Appraisal: A Survey”. In:

Pulkkinen, Urho and Kaisa Simola (2000). *An Expert Panel Approach to Support Risk-Informed Decision Making*. Tech. rep. STUK-YTO-TR 170. ISBN 951-712-421-X. Helsinki: STUK - Finland Radiation and Nuclear Safety Authority, p. 16.

Ranger, Nicola and Falk Niehörster (2012). “Deep Uncertainty in Long-Term Hurricane Risk: Scenario Generation and Implications for Future Climate Experiments”. en. In: *Global Environmental Change* 22.3, pp. 703–712. ISSN: 09593780. DOI: 10.1016/j.gloenvcha.2012.03.009.

Risser, Mark D. and Michael F. Wehner (2017). “Attributable Human-Induced Changes in the Likelihood and Magnitude of the Observed Extreme Precipitation during Hurricane Harvey”. en. In: *Geophysical Research Letters* 44.24, pp. 12,457–12,464. ISSN: 1944-8007. DOI: 10.1002/2017GL075888.

Romeijn, Jan-Willem (manuscript). “Pooling, Voting, and Bayesian Updating”. In: DOI: http://www.philos.rug.nl/~romeyn/paper/2015_romeijn_-_pooling_voting_updating.pdf.

Rosenstock, Sarita, Cailin O’Connor, and Justin Brunner (2017). “In Epistemic Networks, Is Less Really More?” In: *Philosophy of Science* 84.2, pp. 234–52.

Rougier, Jonathan (2016). “Ensemble Averaging and Mean Squared Error”. In: *Journal of Climate* 29.24, pp. 8865–8870. ISSN: 0894-8755. DOI: 10.1175/JCLI-D-16-0012.1.

Rudner, Richard (1953). “The Scientist Qua Scientist Makes Value Judgments”. en. In: *Philosophy of Science* 20.1, pp. 1–6. ISSN: 0031-8248, 1539-767X. DOI: 10.1086/287231.

Sabbatelli, Tom (2017). *Catastrophe Modeling - Part 2*. en-US.

Sabbatelli, Tom and Jeff Waters (2015). *We’re Still All Wondering – Where Have All The Hurricanes Gone?*

Savage, L J (1971). “Elicitation of Personal Probabilities and Expectations”. en. In: *Journal of the American Statistical Association* 66.336, pp. 783–801.

Shome, Nilesh, Mohsen Rahnama, Steve Jewson, and Paul Wilson (2018). “Quantifying Model Uncertainty and Risk”. en. In: *Risk Modeling for Hazards and Disasters*. Elsevier, pp. 3–46. ISBN: 978-0-12-804071-3. DOI: 10.1016/B978-0-12-804071-3.00001-X.

Simion, Mona, Christoph Kelp, and Harmen Ghijsen (2016). “Norms of Belief”. en. In: *Philosophical Issues* 26.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/phis.12077, pp. 374–392. ISSN: 1758-2237. DOI: 10.1111/phis.12077.

Stainforth, David A., M.R. Allen, E.R. Tredger, and Leonard Smith (2007a). "Confidence, Uncertainty and Decision-Support Relevance in Climate Predictions". In: *Philosophical Transactions of the Royal Society A* 365.June, pp. 2145–2161.

Stainforth, David A., T. E Downing, R. Washington, A. Lopez, and M. New (2007b). "Issues in the Interpretation of Climate Model Ensembles to Inform Decisions". In: *Philosophical Transactions of the Royal Society A* 365.1857, pp. 2163–2177.

Steele, Katie (2012). "Testimony as Evidence: More Problems for Linear Pooling". en. In: *Journal of Philosophical Logic* 41.6, pp. 983–999. ISSN: 0022-3611, 1573-0433. DOI: 10.1007/s10992-012-9227-5.

Stefánsson, H. Orri and Katie Steele (ms). *Belief Revision for Growing Awareness*.

Strachan, Jane, Pier Luigi Vidale, Kevin Hodges, Malcolm Roberts, and Marie-Estelle Demory (2012). "Investigating Global Tropical Cyclone Activity with a Hierarchy of AGCMs: The Role of Model Resolution". In: *Journal of Climate* 26.1, pp. 133–152. ISSN: 0894-8755. DOI: 10.1175/JCLI-D-12-00012.1.

Suppes, Patrick (1969). "A Comparison of the Meaning and Uses of Models in Mathematics and the Empirical Science". In: *Studies in the Methodology and Foundations of Science*. Ed. by Patrick Suppes. Dordrecht: Reidel, pp. 10–23.

Tebaldi, C. and R. Knutti (2007). "The Use of the Multi-Model Ensemble in Probabilistic Climate Projections". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 365.1857, pp. 2053–2075. ISSN: 1364-503X. DOI: 10.1098/rsta.2007.2076.

Teller, Paul (2001). "Twilight of the Perfect Model Model". en. In: *Erkenntnis* 55.3, pp. 393–415.

Thoma, Johanna (2019). "Decision Theory". In: *The Open Handbook of Formal Epistemology*. Ed. by Richard Pettigrew and Jonathan Weisberg. PhilPapers Foundation.

Titelbaum, M. G. (2012). *Quitting Certainties: A Bayesian Framework Modeling Degrees of Belief*. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780199658305.001.0001.

Titelbaum, Michael G. (forthcoming). *Fundamentals of Bayesian Epistemology*.

Wagner, Carl (2002). "Probability Kinematics and Commutativity". In: *Philosophy of Science* 69.2, pp. 266–78.

Weisberg, Michael (2007a). "Three Kinds of Idealization". In: *The Journal of Philosophy* 104.12, pp. 639–659. ISSN: 0022-362X.

— (2007b). "Who Is a Modeler?" en. In: *The British Journal for the Philosophy of Science* 58.2.

Weisberg, Michael (2013). *Simulation and Similarity: Using Models to Understand the World*. Oxford: Oxford University Press.

Williamson, Timothy (2006). "Must Do Better". en. In: *Truth and Realism*. Ed. by Patrick Greenough and Michael P. Lynch. OCLC: ocm65201292. Oxford : New York: Clarendon Press ; Oxford University Press, pp. 177–188. ISBN: 978-0-19-928888-5 978-0-19-928887-8.

— (2017). "Model-Building in Philosophy". In: *Philosophy's Future: The Problem of Philosophical Progress*. Ed. by Russell Blackford and Damien Broderick. Oxford: Wiley.

Wimsatt, William C. (2007). *Re-Engineering Philosophy for Limited Beings*. Cambridge, MA: Harvard University Press.

Winsberg, Eric (2018). *Philosophy and Climate Science*. Cambridge: Cambridge University Press.

Woudenberg, F (1991). "An Evaluation of Delphi". In: *Technological Forecasting and Social Change* 40.2, pp. 131–150.

Zollman, Kevin J. S. (2010). "The Epistemic Benefit of Transient Diversity". In: *Erkenntnis* 72.1, pp. 17–35. ISSN: 0165-0106.

van Fraassen, Bas (1980). *The Scientific Image*. Oxford University Press.

— (1981). "A Problem for Relative Information Minimizers in Probability Kinematics". en. In: *The British Journal for the Philosophy of Science* 32.4, pp. 375–379. ISSN: 0007-0882. DOI: 10.1093/bjps/32.4.375.