Four essays in mathematical philosophy

Rafiee Rad, S.

Publication date:
2014

Document Version
Publisher's PDF, also known as Version of record

Four Essays in Mathematical Philosophy

Proefschrift

ter verkrijging van de graad van doctor
aan Tilburg University
op gezag van de rector magnificus,
prof. dr. Ph. Eijlander,
in het openbaar te verdedigen ten overstaan van een
door het college voor promoties aangewezen commissie
in de aula van de Universiteit

op maandag 29 september 2014 om 10.15 uur

door

Soroush Rafiee Rad

geboren op 22 juli 1981 te Tehran, Iran

Promotiecommissie

Promotor: prof. dr. S. Hartmann

Overige leden van de Promotiecommissie:

prof. dr. L. Bovens
prof. dr. I. Douven
prof. dr. R. Parikh
dr. S. Smets
prof. dr. J. Sprenger

Acknowledgments

I would like to thank my parents for their constant support, encouragement and persuasion, for nurturing my curiosity and teaching me the value of knowledge, and for providing me with a lifetime of opportunities that have led to this point.

I would like to express special gratitude to my supervisor, Prof. Stephan Hartmann, for his patience, inspiration, contribution and guidance, without which this work would not have been possible. I am grateful to him for teaching me how to think as a philosopher and for four fruitful years of intellectual stimulation.

I would like to thank the members of my thesis committee for many constructive comments and suggestions that have greatly improved the work presented in this thesis.

I am grateful to my brother Siavash for his help, support and encouragement throughout all the years of my studies, as well as for his help in proof reading major parts of this thesis and for many valuable discussions that have led to many improvements in this work.

I am also grateful to my friend Karim Thebault for very useful comments that significantly improved the presentation of this thesis.

Finally I would like to thanks my friend Arash Eshghi, for endless hours of discussion, for many years of intellectual stimulation and for his ever increasing love for philosophy that has been a significant catalyst to my philosophical curiosity for many years. For these I am grateful to him, and the work in this thesis is directly or indirectly indebted to him.

ABSTRACT.

Scientific philosophy is a recent but rapidly growing approach to investigate a wide range of philosophical problems. This approach advocates the employment of scientific methodologies, including mathematical and computational methods, in philosophical investigations. In this thesis we will present four case studies in scientific philosophy, using both mathematical/logical formalisations and computational simulations. We will investigate problems from different philosophical disciplines aiming to show how the formal and computational methods can be beneficial to a wide range of philosophical investigations. We shall study a probabilistic approach to para-consistency and reasoning from conflicting information, learning indicative conditionals, modelling rational deliberation and an investigation of the anchoring effect in deliberations. All these are long standing problems in philosophy that have attracted a lot of interest, in particular, in recent years. Thus, in this thesis, we hope to contribute to the growing literature in scientific philosophy and further motivate its extensive domain of application.

Contents

Chapter 1

Introduction

Scientific philosophy is a relatively new approach to philosophical enquiry that amidst, sometimes strongly voiced, oppositions and supports is increasingly gaining popularity in the recent years. According to Leitgeb (Leitgeb 2013) there are at least three ways to understand scientific philosophy.

The first view goes back to the ideas of the Vienna and the Berlin circles and considers philosophy as a discipline employed in the service of science. On this view, the role of philosophy is to study and analyse science on the meta-level. The goal of philosophy will then be to refine and improve scientific language and to enhance the logic where needed and possible. The study of scientific fields on the meta-level clarifies and improves our understanding of the concepts that play a part in the corresponding scientific field and their interrelation. In doing so, philosophy contributes to the development of an appropriate object language with which scientists work and the adoption of the correct logical structures and reasoning mechanisms suitable for different scientific theories. The contribution of philosophy in this view, as advocated by Michael Friedman for example, is most visible in Kuhnian paradigm shifts when a new scientific theory replaces an old one. At such revolutionary stages, philosophy plays a crucial role in development of proper scientific language and is the force that ensures the scientific theory change remains confined within the boundaries of rationality.

The second view sees philosophical studies as part of the scientific endeavour. This account, advocated amongst others by Quine, roots in Naturalism and considers natural sciences our best medium to access the truth and so the most viable approach to philosophical studies. According to this view, the role of philosophy is to analyse and shed light on the foundational issues in scientific disciplines and the philosophical studies should be carried out along the same methodological lines, in the same language as in the respective scientific

fields and by employment and use of scientific theories and their results. Quine's naturalised epistemology, for example, emphasises the close connection between epistemology and natural sciences and the necessity to take into account the results from the study of human reasoning when dealing with problems in epistemology. In his view, our knowledge can be captured mainly in our language use, the observable characteristics of which allows its study in the same manner as other scientific inquiries. This account closely connects the study of epistemology with research in cognitive psychology and argues in favour of articulating problems and arguments in epistemology using the language and concepts developed in psychology. As pointed out by Leitgeb (Leitgeb 2013), philosophy of mathematics and set theory as well as studies in metaphysics and physics, for example, bear the same connection and in this conception of scientific philosophy should be carried out in the same languages respectively and deal with the same issues and concerns. Thus this view urges (or in more radical approaches, necessitates) the employment and use of scientific theories and the results of investigations in relevant scientific disciplines for proper and fruitful study of philosophical questions. This approach has been taken up in more recent years by philosophers such as Penelope Maddy in her book "Second Philosophy" (Maddy 2009), and James Ladyman and Don Ross in their account of naturalistic metaphysics in "Every Thing Must Go: Metaphysics Naturalized" (Ladyman & Ross 2009) who have argued for naturalisations of parts of philosophy.

Finally the third view, understands scientific philosophy as the philosophy carried out using scientific methods. This is the account of scientific philosophy that this thesis will fall into. As emphasised by Leitgeb, this account is in no conflict with viewing philosophy as an independent discipline that is "not necessarily being pursued, whether on the meta-level or on the object level, with the aim of facilitating scientific progress" (Leitgeb 2013). Thus with this account of scientific philosophy, philosophers are no more required (at least not necessarily) to be concerned with problems and notions raised in scientific theories and philosophy is an autonomous discipline with its own concepts and questions. This view is thus consistent with philosophy as a discipline that is pursued for the sake of understanding issues and concerns that stem from motivations other than scientific progress and that have engaged our forefathers since antiquity; issues such as truth, knowledge, existence, morality and ethics.

What, then, are the scientific methods that that can be employed in philosophical studies? As hard as it may be to clearly characterise what can or cannot be considered as scientific methods useful for philosophy, there are some obvious candidates to start with. In (Leitgeb 2013), for instance, Leitgeb points to three examples, namely, mathematical, computational and experimental methods. First are the mathematical methods. Mathematical methods have been in use in scientific studies for as long as such studies have been carried out. Their

employment in philosophy is also old news. Application and study of mathematical logic in philosophical studies, for example, dates back to works of Aristotle. Leibniz work on metaphysics is another example in point. The link has grown stronger in the more contemporary studies in philosophy by introduction of inductive logic, temporal logics, epistemic logics, dynamic logics and the like. The connection, however, does not end with mathematical logic and includes a wide range of mathematical disciplines such as probability theory, game theory, discrete mathematics, etc. The interplay of philosophy and mathematics in the investigation of mechanisms for reasoning, in particular, has brought about thriving research programs that have continued throughout the 20th century and is continuing to this date. An impressive collection of works in the literature, commonly referred to as Formal Philosophy is witness to this thriving dynamics. The formal Epistemology movement and the birth and formulation of Bayesian Epistemology are examples in point.

Second are the computational methods. This is, in short, application of computational algorithms, computer simulations and the resulting numerical analysis in the study of philosophical problems and in support of philosophical claims and arguments. Computational models, for example, have been used for more than two decades to study both the process of scientific discovery and the process of evaluating scientific theories. The BACON project, a pioneering project in this regard, for instance, was developed by Pat Langley, Herbert Simon and their colleagues (Langley et. al. 1987) as model for deriving mathematical laws from numerical data. Other examples are the KEKADA program developed by Kulkarni and Simon (Kulkarni & Simon 1988) and Paul Thagard "Computational Philosophy of Science" (Thagard 1993), that introduces a computational model for problem solving which he uses to study issues in philosophy of science such as hypothesis formation and theory justification amongst others. The Structure Mapping Engine developed by Falkenhainer, Forbus, and Gentner (Falkenhainer et. al. 1989), the Analogical Constraint Mapping Engine introduced by Holyoak and Thagard (Holyoak & Thagard 1989; Holyoak & Thagard 1995) for modelling anological reasoning and the ECHO program developed by Thagard to model theory evaluation in science are other examples of computational approaches as is the Fitelson's and Zelta's (Fitelson & Zalta 2007) work on computational metaphysics.

More recently, there has been many examples of the application of computer simulations, in particular formal epistemology and in philosophy of social sciences. This is, at least, partly because these simulations prove very useful in studying the group dynamics and the emergence and evolution of phenomena in social interactions. Many instances of (emergence or dissolution of) social phenomena or patterns in groups depend on the group topology and the connections between individuals in the group. This makes the analytical study of such

instances difficult in the sense that many different cases have to be considered separately. Computer simulations allow study of diverse large sample spaces to identify common trends that that can be studied uniformly among different cases. Not only are computer simulations increasingly used in philosophical studies, but their application has become significant enough to prompt philosophical studies into simulations themselves. These studies, for instance "Computer Simulations and the Changing Face of Scientific Experimentation" by Juan Duran and Eckhart Arnold (Duran & Arnold 2013), are aimed to better understand the type of inference that is possible on the basis of computer simulations and the criteria that constitute a conceptually adequate simulation.

Third kind of scientific methods, are the experimental methods. Experimental philosophy uses empirical data gathered through surveys, interviews and designed experiments to derive clues with regard to philosophical questions. The issue of using such methods is the subject of intense debate among philosophers. Although the empirical studies appear to be illuminating and insightful for philosophical studies in many instances, in particular, investigations in philosophy of mind, philosophy of language, morality and the like, for example in the works of Natalie Gold, Regina Rini, Stephen Stich, Shen-yi Liao among others, there has been serious criticisms raised in opposition to it. One immediate reason is that, as opposed to computational and mathematical methods, empirical results does not provide a priori knowledge or justification. Another point of criticism is that people's intuitions, which are the main outcome of experimental studies, cannot be considered as evidence for philosophical studies. Thus, the use of experimental methods is more controversial than the formal and computational methods and we shall not be concerned with them in this thesis.

There are many reasons as to why the application of formal methods is useful in philosophy. An important contribution of mathematical methods is the explication of philosophical concepts. This is the development of new concepts that can extend existing concepts in the sense that the new concept coincides with the old in standard and clear cases while improving on it in "exactness, fruitfulness and simplicity", see (Leitgeb 2013), in the more problematic or fuzzy cases. In (Leitgeb 2013) Leitgeb argues that not only are the mathematical formulations useful for the process of explicating philosophical concepts but they are in many cases necessary. Tarski's explication of truth, Carnap's explication of confirmation of hypotheses by evidence, and Adams' explication of the acceptability of conditionals are examples of such cases. Tarski's work is built on second order logic and set theory while Carnap's and Adams' require the theory of subjective probability. In addition, mathematical definitions and formulation, where possible, can make philosophical concepts more precise and immune to divergence of interpretations. Similarly, precision and exactness of mathematical proofs can be carried into the philosophical arguments that are formulated in mathematical

terms. Using mathematical language and formal methods not only makes the philosophical arguments more precise but are also needed to represent the more complex arguments correctly and more understandably. The same way formal models are useful in presenting arguments that are not necessarily proofs in the sense of showing that the truth of assumptions entail the truth of the conclusions but rather inductively strong in the sense that the conclusions are at least as likely as the assumptions. Hartmann and Bovens use of Bayesian networks in (Bovens & Hartmann 2003) in support of claims about confirmation, testimony, etc. have such characteristic. What is more, mathematical formulations will make all the relevant assumptions and prerequisites explicit and again prevent the multitude of interpretations. They clarify links to scientific theories which might in turn introduce interesting new philosophical concepts or questions and even point to enological cases in different areas of philosophy.

The benefits of such formal approaches, however, can be best evaluated by looking at the contribution of the application of such methods to the philosophical literature. The role of mathematical formulation is robustness and rigidity of Bayesian epistemology, theories of truth, studies of rationality and belief revision and investigations in social epistemology and collective rationality and decision making needs no reminder. It seems, however, important to emphasise that the formal approach to philosophy is not a reductionist view. The aim is not to reduce philosophical studies to mathematics or any other scientific discipline but rather to use mathematical, computational and scientific results to the benefit of philosophical investigations where such applications are possible. This is to acknowledge that there may very well remain many areas, concepts and questions in philosophy where it is not possible (or not yet possible) to take a formal stand. But where the application of such methods has been possible, the input and benefit of such applications to the literature, to which we also hope to contribute in this thesis, is undeniable.

What Follows....

In what follows we will present four studies in scientific philosophy in the third sense above, using mathematical and computational methods (and their combination). The studies are carried out in the framework of formal and social epistemology and the first two deal with problems of reasoning for individual rational agents and the second two are concerned with issues of collective decision making. All problems that we shall visit in this thesis have been of long standing interest to philosophers and the subject of philosophical debate and investigation for quite some time. The goal here is to demonstrate how the application of formal tools can help to settle or further elaborate these problems.

We start with the problem of para-consistent reasoning. This has attracted the theoretical interest of logicians and philosophers for a long time and sev-

eral approaches have been proposed and studied in the literature, see for example, (da Costa 1974; da Costa 1989; da Costa 1998) (Priest 1979; Priest 1987; Priest 1989). Besides the purely theoretical interest, however, working with inconsistencies is of great importance in the study of practical reasoning. Our approach to para-consistency arises from the studies of probabilistic consequence relations in (Knight 2002), (Paris 2004) and (Picado-Muino 2008). We advocate the idea that the inconsistency in an agent's evidence should be identified with the uncertainty that it will induce in the agent's knowledge. In this sense, reasoning with inconsistent information is essentially reduced to uncertain reasoning. We will proceed in our investigation in three steps. First, we will present the formal machinery to bridge between an inconsistent set and its uncertain consistent reduction. Next, we will extend the approach developed in (Paris 2004) and (Paris et. al. 2008) for defining a probabilistic consequence relation, to first order languages. The probabilistic consequence relation will then provide the formal logical system for reasoning with these (consistent) uncertain knowledge sets. Finally we will briefly discuss some immediate generalisations of this approach.

Almost all current models of belief revision assign a higher degree of reliability to the new information than what is already in the belief set. The approach presented in this study allows us to assign different degrees of trust to the newly received information not only with respect to the current knowledge set as a whole but also with respect to each individual statement in that set. This will, thus, allow a fine graded analysis of the inconsistency in relation to the current knowledge set and the new information.

The second study on the individual aspects of reasoning concerns the indicative conditionals. The issue has been studied meticulously in the works of Arlo-Costa, Lewis, Stalnaker, van Fraassen and Douven among others, (Arlo-Costa1990), (Douven 2012), (Douven & Romeijn 2012), (van Fraassen 1981), (Stalnaker 1968), (Lewis 1976), and several approaches have been proposed and studied in the literature. These include identifying the indicative conditional with the corresponding material conditional, working with the Stalnaker account using *imaging* as proposed by David Lewis, updating with Adam's conditioning rule or using information theoretic updating procedures such as Kullback-Leibler (KL) distance minimisation[1].

All these proposals, however, have been criticised by means of counter examples and despite the extensive effort spent on the issue, a general Bayesian account of updating on conditionals is still missing from the literature. The essence of these counter examples deals with the unintuitive effect of updating procedure on the probability of the antecedent of the conditional. To be more precise, they are all concerned with how the posterior probability of the antecedent compares to its prior probability as the result of the updating procedure as opposed to how

[1]This approach was, to our knowledge, first studied by van Fraassen.

they should compare intuitively (see, (Douven 2012)).

We shall address this problem using existing links to measure theory and the Kullback-Leibler Distance minimisation procedure, the theory of casual structures and their relation to conditionals. Our proposal is the implementation of the KL distance minimisation in a slightly richer setting. To be more precise, we will show that the KL distance minimisation will provide the intuitively expected results if applied in a setting where all the relevant variables in the scenario are fixed and the complete causal structure of the problem is identified. In this setting, one will start with the prior belief function induced by the causal structure and the indicative conditional will give a constraint on the posterior probabilities. The posterior probability function will then be chosen as to minimises the KL distance to the priors. We shall revisit all the examples given by Douven and his co-authors as well as the Judy Benjamin example and will show that the above proposal will give the results that one intuitively expects in all proposed scenarios. These two studies are concerned with aspects of reasoning and the dynamics of belief in individual agents while the next two will deal with belief dynamics in a collective setting.

Our third study deals with the investigation of rational deliberation in groups. The expansion and multitude of different social networks, to which almost each and every member of the society is subscribed in the modern day, is rapidly increasing the number of essentially social judgments. The majority of decisions are no longer made by individual agents but rather by the social networks to which they belong and as such, are inevitably subject to influence and revision as they evolve from personal judgments into a collective decisions. This evolution takes place, to a major part, in the course of agents social interactions and communications. Deliberation is an important example of such social interaction and communications and the normative investigation of mechanisms that govern the flow of deliberative processes and the dynamics of belief change towards the group consensus, can be instrumental in devising interaction protocols that facilitate dissemination of some independent and inter-subjective truth, when such is definable. Our goal in this study is to contribute to this investigation.

Groups can proceed to make a collective decision either by aggregating individual judgments such as in voting scenarios or can deliberate on the issue until they reach a consensus where all the group members manifest the same individual judgment. There are also two views on how to evaluate methods for collective decision making. From the proceduralist point of view, a decision making procedure should be judged on the basis of its procedural characteristics only without any reference to the epistemic nature of the outcomes. On the other hand is the epistemic view that evaluates a decision making process by the epistemic values of its outcomes without taking account of the procedural considerations. In case of democratic decision making, the distinction between the two conceptions can

be formulated in terms of the characteristics expected from the outcome: Is it the fairness or the correctness that constitute our main concern?

In this regard, it is not hard to justify that the deliberation presents an obvious procedural advantage to voting. The prospect of achieving a collective consensus, on which all the group members agree, eliminates the necessity of a compromise that is inevitable in voting scenarios and makes the deliberation an ideal approach from procedural point of view. A question of interest is then to ask how the two process compare epistemically. We shall, thus, emphasise our concern on the epistemic nature of the deliberative process and the epistemic comparison between deliberation and voting.

To this end, we shall first introduce a Bayesian model that is built on the basis of two attributes of the decision makers; the *first order reliability*, that is the reliability of each individual to give the correct answer to the problem under deliberation and, the *second order reliability*, that represents each individual's ability to assess the first order reliabilities of her group members. We will then use a combination of mathematical formulations and computer simulations to investigate its truth tracking properties and give a comparison between the epistemic properties of this deliberation model and that of the majority voting.

The fourth, and final, study in this thesis concerns the investigation of the anchoring effect in deliberations. As we emphasise the social aspects of reasoning and multi-agent decision making, there are certain socio-psychological considerations that become relevant to the dynamics of belief change. It is not at any rate surprising that the epistemic and procedural advantages that arise from the interaction and communication between decision makers is also accompanied with certain biases and undesirable factors. Some of these factors, such as the emergence of pluralistic ignorance in groups, have been formally studied in some recent works but there is still a gap in the literature of Bayesian Epistemology in this regard. One of these biases which has been extensively discussed in cognitive psychology, but is surprisingly missing from the formal studies in collective decision making, is the *anchoring effect*.

Anchoring is the common human tendency to rely too heavily on one piece of information in the process of decision making. The effect occurs in a deliberation process when the outcome of the deliberation depends on the order in which different group members present their opinions. More specifically, the studies in cognitive psychology suggests that the group member who speaks first will usually have the highest effect on the final decision of the group, where she is said to have anchored the deliberation. The effect is usually attributed to what is known as the bounded rationality. This refers to cognitive limitations of the decision makers including short attention span, memory loss, deterioration of cognitive ability by fatigue, etc. The question that we will be interested in is whether this bias arise as the result of cognitive limitations only, or can it also

appear in groups of fully rational agents.

In this final study, we will first present a model of rational deliberation with incremental updating procedure as a modification of the Lehrer-Wagner model. We will then use this model to study the path dependence in the deliberation and will show that the anchoring bias can emerge in fully rational groups without any cognitive limitation and merely as a result of such updating procedures.

Chapter 2

Reasoning From Conflicting Information: A First Order Account

2.1 Introduction

The treatment of inconsistencies is a long standing issue for mathematical logic. The process of reasoning in the classical logic has been devised with strong built-in consistency assumptions and it follows that the full force of classical entailment relation is too strong for reasoning with inconsistencies. Although limiting the scope of logical inference to only consistent domains fits well with the spirit of what one requires from reasoning in mathematical contexts, there are many aspects of reasoning where it does not. In particular, we have the case when the context of the reasoning is not assumed to represent some factual property of a structure nor objective facts concerning the real state of things but some not-necessarily-certain information or approximations regarding those facts.

There are different motivations for the development of logics that can accommodate inconsistencies and there have been several attempts in the literature to do so. The main difference between these motivations arise from the way that the inconsistent evidence is interpreted. One motivation stems from adopting the philosophical position of dialetheism, best advocated by Graham Priest for example. This position is characterised by submitting to the thesis that there are sentences which are true and false simultaneously, see for example (Priest 1979; Priest 1987; Priest 1989). One approach to deal with inconsistencies in this view is to adopt a three valued logic with truth values $\{0, 1, \{0, 1\}\}$,

for example, with truth value $\{0, 1\}$ for the sentences that are assumed to be both true and false.

Other motivations can arise from more pragmatic reasons which deal with reasoning in non-ideal contexts. Here the inconsistencies are interpreted as a property of the information and are taken to be anomalies that point out errors or shortcomings of the reasoners' information (or maybe communication channels). The approaches that arise from this latter motivation, primarily, try to deal with the inconsistent sets by reducing the inference to consistent reasoning. This is done either by defining the logical consequences of such sets on the basis of their maximal consistent subsets as is the case for da Costa's para-consistent logics, (da Costa 1974; da Costa 1989; da Costa 1998), or by first revising the inconsistent sets to consistent ones. For example one might define the set of logical consequences of a possibly inconsistent set Γ as the union (or intersection) of the sets of logical consequences of its maximal consistent subsets. Or one might choose to apply some belief revision process to first arrive at a consistent information set Γ', as in AGM belief revision process for instance, (Alchourron et. al.1985), and make the reasoning on the basis of this consistent set. The idea in an AGM-like belief revision process for example, is that upon receiving some inconsistent information ϕ, one will first retract the part of knowledge base that contradicts this new information and then expands the remaining knowledge set by adding ϕ. The assumption here, however, is that the new information is always more reliable than the old. An assumption which is counterintuitive in many aspects of reasoning. For example when the context of reasoning consists of statements derived from a not-completely-reliable sources or processes that are subject to errors. Even more pointed are cases where the context of reasoning consists of statements accumulated through different sources and processes which do not necessarily agree. This is indeed the case in almost all applications of reasoning outside some mathematical theory. As the information set expands by acquiring new information through possibly conflicting sources and processes, it may very well come to include conflicting and inconsistent evidence without any second order information that warrants discarding parts of these evidence in favour of others. This will void the possibility of using classical entailment (or other variations of it which still get trivialised in the presence of inconsistencies) as it validates any consequence from such an inconsistent set. In this sense having some inconsistency in a (possibly very large) set of evidence will render it completely useless for reasoning. There are many applications of reasoning, however, in which the inconsistencies should intuitively affect the reasoning only partially. As a very simple example, consider sentences ϕ and ψ that share no relation symbols, function symbols or constants (hence have completely irrelevant informational content), then

$$\{\phi, \psi, \neg\phi\} \models \neg\psi$$

many instances of which are counterintuitive. For example, assume a case where ϕ is acquired from a source, say S_1, different from that of $\neg\phi$, say S_2 where both sources agree on ψ. Here the inconsistency of the information regarding ϕ may not provide any reason to affect the reasoning on the part of ψ. This motivates one to fashion inference processes that allow meaningful extraction of information from such sets of information. This is the motivation for what we shall pursue in these Chapter and the aspect of the literature which we hope to contribute to.

The approach presented here, follows the work of Knight, (Knight 2002), Paris, (Paris 2004) and Paris, Picado-Muino and Rosefield, (Paris et. al. 2008), in dealing with the same problem for propositional languages and is motivated by reasoning in non-ideal contexts. This approach lies on the assumption that the inconsistent evidence do not point out the inconsistencies of the reality under investigation but point to an inconsistent valuation of facts. Receiving contradictory information should thus affect such valuations. In this view, receiving some piece of information ϕ while having $\neg\phi$ in our knowledge base has the effect of changing the valuation of ϕ (and thus $\neg\phi$). In case of categorical knowledge (with truth values of zero or one), this means moving from categorical belief in ϕ and $\neg\phi$ to some uncertain valuation of them and in case of probabilistic knowledge this would entail re-evaluation of the probabilities. Our approach is based on two assumptions,

- the inconsistencies are identified with the uncertainty that they induce in the information set
- the information is assumed to be as reliable as possibly allowed by the consistency considerations.

Thus receiving inconsistent information will change the context of reasoning from a categorical one to an uncertain one, which we shall represent by means of probabilities. One can also hope to do so in a way that allows us to limit the pathological effect of inconsistencies to the part of the reasoning relevant to it. To make this clear, suppose as above that one is left, after receiving $\neg\phi$, with the inconsistent knowledge $\{\phi, \psi, \neg\phi\}$ where again ϕ is acquired from source S_1 and $\neg\phi$ from source S_2 while both sources agree on ψ. This inconsistency is accommodated by changing the categorical belief in ϕ and $\neg\phi$ to uncertain one by assignment of probabilities with the probabilities of ϕ and $\neg\phi$ adding up to 1 but without changing the valuation of ψ as it is irrelevant to the inconsistency.

How the change in the information set induced by the inconsistency is carried out, depends on one's approach to the weighting of the new information with respect to the old information. For example, if we take the new information to be infinitely more reliable than the old, we will end up with the same retraction and expansion process as in the AGM. But as we shall shortly see, one can also devise the change in a manner that allows a wider range of epistemic attitudes

towards the new information in comparison to the old. Since the inconsistencies will reduce our categorical knowledge to probabilistic one, any inference based on such knowledge will essentially be probabilistic. Our goal is to study an entailment relation that allows meaningful inference from such probabilistic knowledge bases. The idea is to investigate a consequence relation that generalises the classical consequence relation from a relation that preserves the truth to one that preserves, or more precisely ensures, some degree of reliability. To this end we will first investigate how to accommodate inconsistencies of evidence in the information set and will then study a probabilistic entailment relation on propositional languages introduced by Knight, (Knight 2002) and further investigated by Paris, (Paris 2004), and Paris, Picado-Muino and Rosefield, (Paris et. al. 2008), for the first order case in order to make inference on the basis of such uncertain knowledge bases.

It is also worth mentioning that one can choose a different route altogether and deal with the inconsistent evidence by adopting a richer language in which the source of information is also coded in the information. Thus, for example, ϕ received from source S_1 is replaced by $(\phi)^1$ to the effect that "according to S_1, ϕ". In this approach receiving ϕ^1 (according to S_1, ϕ) and $(\neg\phi)^2$ (according to S_2, $\neg\phi$) pose no contradiction any more while contradictory information from the same source has the effect of reducing the reliability of the source. The evaluation of information is carried out by weighting them with the reliability of the sources. As it would be immediately clear however, this approach will be equivalent to ours. The simplest case we will discuss corresponds to receiving information from equally reliable sources. The case of prioritised evidence corresponds to receiving information from sources with different reliabilities. Our approach, however, has the advantage of avoiding unnecessary complication of the language.

The rest of this chapter is organised as follows. In Section 2.2 we will investigate a revision process for reducing inconsistent information sets to (probabilistically consistent) uncertain ones. We will investigate revision of categorical information in Section 2.2.1, probabilistic information in Section 2.2.2 and prioritised information in Section 2.2.3. In the Section 2.3 we will investigate a probabilistic entailment relation that allows meaningful inferences on possibly inconsistent sets. We shall give an analysis of this entailment relation in the first order logic in Section 2.7.1 and will next investigate a generalisation that allows different epistemic status for individual sentences in the knowledge set and thus providing the setting to limit the effect of inconsistency to only part of the reasoning in Section 2.4. In Section 2.5 we will connect this entailment relation to reasoning from conflicting information. Finally the Appendix contains some of the longer and more involved proofs.

2.1.1 Preliminaries and Notation

Throughout these chapter we will work with a first order language $\mathcal{L}$ with finitely many relation symbols, no function symbols and countably many constant symbols $a_1, a_2, a_3, ...$ Furthermore we assume that these individuals exhaust the universe. This means in particular that we have a name for every element in our universe. Thus a model is a structure M for the language $\mathcal{L}$ with domain $|M| = \{ a_i \,|\, i = 1, 2, ...\}$ where every constant symbol is interpreted as itself.Let $R\mathcal{L}, S\mathcal{L}$ denote the set of relation and the set of sentences of $\mathcal{L}$ respectively.

Definition 2.1.1 *We shall call $w : S\mathcal{L} \to [0\,,\,1]$ a probability function if for every $\phi, \psi, \exists x \psi(x) \in S\mathcal{L}$,*

- *P1. If $\vDash \phi$ then $w(\phi) = 1$.*
- *P2. $w(\phi \vee \psi) = w(\phi) + w(\psi) - w(\phi \wedge \psi)$.*
- *P3. $w(\exists x \psi(x)) = \lim_{n\to\infty} w(\bigvee_{i=1}^{n} \psi(a_i))$.*

Let $\mathcal{L}$ be a propositional language with propositional variables $p_1, p_2, ..., p_n$. By *atoms* of $\mathcal{L}$ we mean the set of sentences $\{ \alpha_i \,|\, i = 1, ...J\}$, $J = 2^n$ of the form

$$\pm p_1 \wedge \pm p_2 \wedge ... \wedge \pm p_n.$$

By disjunctive normal form theorem, for every sentence $\phi \in S\mathcal{L}$ there is unique set $\Gamma_\phi \subseteq \{ \alpha_i | \, i = 1, ..., J \}$ such that

$$\vDash \phi \leftrightarrow \bigvee_{\alpha_i \in \Gamma_\phi} \alpha_i.$$

It can be easily checked that $\Gamma_\phi = \{ \alpha_j \,|\, \alpha_j \vDash \phi \}$.

Thus if $w : S\mathcal{L} \to [0\,,\,1]$ is a probability function then

$$w(\phi) = w(\bigvee_{\alpha_i \vDash \phi} \alpha_i) = \sum_{\alpha_i \vDash \phi} w(\alpha_i)$$

as the α_i's are mutually inconsistent. On the other hand since $\vDash \bigvee_{i=1}^{J} \alpha_i$ we have $\sum_{i=1}^{J} w(\alpha_i) = 1$. So the probability function w will be uniquely determined by its values on the α_i's, that is by the vector

$$< w(\alpha_1), ..., w(\alpha_J) > \in \mathbb{D}^{\mathcal{L}} \quad \text{where} \quad \mathbb{D}^{\mathcal{L}} = \{ \vec{x} \in \mathbb{R}^J \,|\, \vec{x} \geq 0, \sum_{i=1}^{J} x_i = 1 \}.$$

Conversely if $\vec{a} \in \mathbb{D}^{\mathcal{L}}$ we can define a probability function $w' : S\mathcal{L} \to [0, 1]$ such that $< w'(\alpha_1), ..., w'(\alpha_J) >= \vec{a}$ by setting

$$w'(\phi) = \sum_{\alpha_i \models \phi} a_i.$$

This gives a one to one correspondence between the probability functions on $\mathcal{L}$ and the points in $\mathbb{D}^{\mathcal{L}}$. In particular if a knowledge base K is taken to be a *satisfiable* set of linear constraints of the form

$$\sum_{j=1}^{n} a_{ij} w(\phi_j) = b_i, \quad i = 1, 2, ..., m$$

where $\phi_j \in S\mathcal{L}$, $a_{ij}, b_j \in \mathbb{R}$ and w is a probability function, then replacing each $w(\phi_j)$ in K with $\sum_{\alpha_i \models \phi_j} w(\alpha_i)$ and adding the equation $\sum_{i=1}^{J} w(\alpha_i) = 1$ we will get a new set of constraints given in terms of the probability of atoms

$$\sum_{j=1}^{J} a'_{ij} w(\alpha_j) = b_i, \quad i = 1, 2, ..., m$$

$$< w(\alpha_1), ..., w(\alpha_J) > A_K = \vec{b}_K.$$

The situation for first order languages is a bit more complicated. Here the atoms of the language are defined as the set of formulas

$$\bigwedge_{\substack{R\ j-ary \\ R \in R\mathcal{L}, j \in N^+}} \pm R(x_{i_1}, ..., x_{i_j}).$$

In the case of first order languages, what plays the role similar to the atoms for a propositional language, are the state descriptions.

Definition 2.1.2 *Let $\mathcal{L}$ be a first order language with the set of relation symbols $R\mathcal{L}$ and let $\mathcal{L}^{(k)}$ be a sub-language of $\mathcal{L}$ with only finitely many constant symbols $a_1, ..., a_k$. The* state descriptions *of $\mathcal{L}^{(k)}$ are the sentences $\Theta_1^{(k)}, ..., \Theta_{n_k}^{(k)}$ which enumerate all the sentences of the form*

$$\bigwedge_{\substack{i_1, ..., i_j \leq k \\ R\ j-ary \\ R \in R\mathcal{L}, j \in N^+}} \pm R(a_{i_1}, ..., a_{i_j}).$$

The following theorem, due to Gaifman, provides a similar result, to that we had above, for the case of a first order language $\mathcal{L}$. Let $QFS\mathcal{L}$ be the set of quantifier free sentences of $\mathcal{L}$:

Theorem 2.1.3 *Let $v : QFS\mathcal{L} \to [0,1]$ satisfy P1 and P2 for $\phi, \psi \in QFS\mathcal{L}$. Then v has a unique extension $w : S\mathcal{L} \to [0,1]$ that satisfies P1, P2 and P3. In particular if $w : S\mathcal{L} \to [0,1]$ satisfies P1, P2 and P3 then w is uniquely determined by its restriction to $QFS\mathcal{L}$.*

For $\phi \in QFS\mathcal{L}$ let k be an upper bound on the i such that a_i appears in ϕ. Then ϕ can be thought of as being from the propositional language $\mathcal{L}^{(k)}$ with propositional variables $R(a_{i_1}, ..., a_{i_j})$ for $i_1, ..., i_j \leq k$, $R \in R\mathcal{L}$ and R $j-ary$. Then the sentences $\Theta_i^{(k)}$ will be the atoms of $\mathcal{L}^{(K)}$ and

$$\phi \leftrightarrow \bigvee_{\Theta_i^{(k)} \models \phi} \Theta_i^{(k)} \quad so \quad w(\phi) = \sum_{\Theta_i^{(k)} \models \phi} w(\Theta_i^{(k)}).$$

Thus to determine the value $w(\phi)$ we only need to determine the values $w(\Theta_i^{(k)})$ and to require

- $w(\Theta_i^{(k)}) \geq 0$ and $\sum_{i=1}^{n_k} w(\Theta_i^{(k)}) = 1$.
- $w(\Theta_i^{(k)}) = \sum_{\Theta_j^{(k+1)} \models \Theta_i^{(k)}} w(\Theta_j^{(k+1)})$,

to ensure that w satisfies P1 and P2. Using this we will limit ourselves to only dealing with $QFS\mathcal{L}$.

2.2 Revising Inconsistent Evidence

2.2.1 Revising Inconsistent Categorical Evidence

We will first investigate the question of how to revise the evidence sets B when receiving inconsistent information; that is when receiving a new piece of information θ where $B \cup \{\theta\} \models \bot$. As mentioned above using an AGM like revision process assumes that new information is always more reliable than the old information. An assumption that is problematic in many contexts of reasoning. Our aim here is to devise a revision process that relaxes this assumption. In our first attempt we assume the same epistemic status for the new information as for any of the statements currently in the evidence set B. We shall relax this assumption in the next sections to allow for a more detailed analysis of the evidence and to take into account the degree of reliability for each individual piece of evidence.

Assume for start that the agent is in possession of a consistent set $B = \{\phi_1, \ldots, \phi_n\}$. We start by assuming categorical information only and will extend our setting to allow for probabilistic evidence later. Suppose that some new piece of information, say θ, is received by the agent where $B \cup \{\theta\}$ is inconsistent. Following our initial intuition this inconsistency will induce uncertainty in the

agent's belief and thus results in moving to some probabilistic belief set B' which is intended to represent a probabilistically consistent reduction of $B \cup \{\theta\}$, i.e., a set B' consisting of probabilistic statements of the form $w(\phi) = p$ for $\phi \in B \cup \{\theta\}$.

Definition 2.2.1 *(Knight 2002) For a set of sentences $\Gamma \subset S\mathcal{L}$, the maximal consistency of Γ, denoted by $mc(\Gamma)$ is defined as*

$$mc(\Gamma) = max\{\eta \,|\, \Gamma \text{ is } \eta \text{ consistent}\} =$$

$$max\{\eta \,|\, \text{there is a probability function } w \text{ on } S\mathcal{L} \text{ such that } w(\phi) \geq \eta \text{ for all } \phi \in \Gamma\}$$

Lemma 2.2.2 *Let $\Gamma = \{\phi_1, \ldots, \phi_n\} \subset S\mathcal{L}$ with $mc(\Gamma) = \eta$. Then there is a fixed subset of Γ, say Γ_1 such that for every probability function w on $S\mathcal{L}$, if $w(\phi) \geq \eta$ for all $\phi \in \Gamma$ then $w(\phi) = \eta$ for all $\phi \in \Gamma_1$.*

Proof Suppose not, then for every $\psi \in \Gamma$ there is a probability function w_ψ (not necessarily distinct) such that $w_\psi(\phi) \geq \eta$ for all $\phi \in \Gamma$ and $w_\psi(\psi) > \eta$. Let

$$w = 1/n \sum_{\psi \in \Gamma} w_\psi$$

then for every $\phi \in \Gamma$ we have

$$w(\phi) = 1/n \sum_{\psi \in \Gamma} w_\psi(\phi) > \eta$$

since every $w_\psi(\phi) \geq \eta$, $\psi \neq \phi$ and $w_\phi(\phi) > \eta$. This is a contradiction with $mc(\Gamma) = \eta$. □

Let $\Gamma \subset S\mathcal{L}$ and let $mc(\Gamma) = \eta_1$ and let Γ_1 as in Lemma 2.2.2. Set

$$\eta_2 = max\{\eta \,|\, w(\psi) \geq \eta \text{ for } \psi \in \Gamma - \Gamma_1$$
$$\text{where } w \text{ is a probability function such that } w(\phi) \geq \eta_1 \text{ for } \phi \in \Gamma\}.$$

With The same argument as in Lemma 2.2.2, one can show that there is a fixed subset $\Gamma_2 \subset \Gamma - \Gamma_1$ such that $w(\theta) = \eta_2$ for $\theta \in \Gamma_2$ and $w(\theta) \geq \eta_2$ for $\theta \in \Gamma - (\Gamma_1 \cup \Gamma_2)$ for every probability function w such that $w(\phi) \geq \eta_1$ for $\phi \in \Gamma$ (so $w(\phi) = \eta_1$ for $\phi \in \Gamma_1$) and $w(\psi) \geq \eta_2$ for $\psi \in \Gamma - \Gamma_1$. Following the same process finitely many times one will be left a partition $\Gamma = \Gamma_1 \cup \Gamma_2 \cup \ldots \cup \Gamma_m$ and values $\eta_1, \ldots, \eta_m$. Then set

$$\vec{mc}(\Gamma) =< \delta_1, \ldots, \delta_n >, \text{ where } \delta_j = \eta_k \iff \phi_j \in \Gamma_k.$$

Intuitively the values given in $\vec{mc}(\Gamma)$ are the highest probabilities that can be assigned to the sentences in Γ consistently. In the sense that there is no probability function that can assign a probability higher than η_1 to all the sentences in

Γ_1 simultaneously and same for η_2 and Γ_2 and so on. In other words if we take $\vec{1} =< 1, \ldots, 1 >$ as an n-vector representing the assignment of reliability 1 to all sentences $\phi_1, \ldots, \phi_n$ (which will be inconsistent if Γ is) then for any probability function w if we set $\vec{w} =< w(\phi_1), \ldots, w(\phi_n) >$, we have

$$d(\vec{1}, \vec{mc}(\Gamma)) \leq d(\vec{1}, \vec{w})$$

thus accounting for $\vec{mc}(\Gamma)$ being the closest we can consistently get to the assumption that all sentences in our knowledge set Γ are correct.

Definition 2.2.3 *Let $B = \{\phi_1, \ldots, \phi_n\} \subset S\mathcal{L}$ be consistent set of sentences and $\phi_{n+1} \in S\mathcal{L}$ be such that $B \cup \{\phi_{n+1}\} \models \bot$, then the revision of B by ϕ_{n+1} is defined as*

$$B' = \{w(\phi_1) = p_1, \ldots, w(\phi_n) = p_n, w(\phi_{n+1}) = p_{n+1}\}$$

where

$$< p_1, \ldots, p_n, p_{n+1} >= \vec{mc}(\{\phi_1, \ldots, \phi_n, \phi_{n+1}\}).$$

Definition 2.2.3 is to capture the idea that the revised belief set is to assign probabilities to the sentences $\phi_1, \ldots, \phi_n, \phi_{n+1}$ that are as close as possible to 1, that is to assign the highest reliability to the information that is consistently possible.

2.2.2 Revising Inconsistent Probabilistic Evidence

Using the revision process described above, one will move, in the presence of inconsistencies, from a set of categorical information to one consisting of probabilistic statements. To use this as a process for iterated revision one needs to define the revision process also on those consisting of probabilistic statements. The latter will be more general and include the categorical information sets by identifying a set $\{\phi_1, \ldots, \phi_n\}$ with the set $\{w(\phi_1) = 1, \ldots, w(\phi_n) = 1\}$.

Notice that in revising the $B = \{\phi_1, \ldots, \phi_n\}$, with a sentence ϕ_{n+1}, the notion of maximal consistency of $B \cup \{\phi_{n+1}\}$ represent an attempt to consistently assign probabilities to these sentences while remaining as close as possible to their prior probabilities (namely, 1). Thus the attempt to assign the highest probabilities consistently possible was essentially an attempt to remain as close as possible to 1. The approach when dealing with probabilistic belief sets in general is going to be the same. We shall try to assign probabilities to these sentences while trying to set the values as close as possible to the prior probabilities, which might not necessarily be 1 any more. To this end we first generalise the notion of maximal consistency for a set Γ. For a set of probabilistic statements, $\Gamma = \{w(\phi_1) = p_1, \ldots, w(\phi_n) = p_n\}$, we say that Γ is inconsistent when there is no probability function W such that $W(\phi_i) = p_i$. In other words when w can not be extended to a probability function.

Definition 2.2.4 *Let $\Gamma = \{w(\phi_1) = p_1, \ldots, w(\phi_n) = p_n\}$ be a (possibly inconsistent) set of probabilistic sentences. The minimal change consistency of Γ, $\vec{mcc}(\Gamma)$, is defined as the n-vector*

$$\vec{q} \in \{< a_1, \ldots, a_n > \mid \text{there is a probability function } W \text{ on } S\mathcal{L} \text{ with } W(\phi_i) = a_i\}$$

for which $d(\vec{q}, \vec{p})$ is minimal, where $\vec{p} =< p_1, \ldots, p_n >$ and d is the Euclidean distance.

Notice that for consistent $\Gamma = \{w(\phi_1) = p_1, \ldots, w(\phi_n) = p_n\}$, the $\vec{mcc}(\Gamma) =< p_1, \ldots, p_n >$. The process of revising a set of probabilistic information$B = \{w(\phi_1) = p_1, \ldots, w(\phi_n) = p_n\}$ with the statement $w(\phi_{n+1}) = p_{n+1}$ is the same as revising categorical information but with $\vec{mcc}(B \cup \{w(\phi_{n+1}) = p_{n+1}\})$ instead of $\vec{mc}(B \cup \{\phi_{n+1}\})$.

Definition 2.2.5 *Let $B = \{w(\phi_1) = p_1, \ldots, w(\phi_n) = p_n\}$, where $\{\phi_1, \ldots, \phi_n\} \subset S\mathcal{L}$ and $\phi_{n+1} \in S\mathcal{L}$ be such that $B \cup \{w(\phi_{n+1}) = p_{n+1}\}$ is probabilistically inconsistent*[1]*, then the revision of B by $w(\phi_{n+1}) = p_{n+1}$ is defined as*

$$B' = \{w(\phi_1) = q_1, \ldots, w(\phi_n) = q_n, w(\phi_{n+1}) = q_{n+1}\}$$

where

$$\vec{q} = \vec{mcc}(B \cup \{w(\phi_{n+1}) = p_{n+1}\}).$$

2.2.3 Revising Prioritised Evidence With Degrees Of Entrenchment

One can immediately notice that in the revision process described above all the sentences in the current belief set, as well as the new information $w(\phi_{n+1})$, are given the same epistemic status, in the sense that one tries to keep them all as close as possible to the prior values. This can be readily relaxed in our setting. One can modify the distance used in the definition of $\vec{mcc}$ to account for a higher degree of reliability or trust in the new information or the old. More generally one can assign degrees of entrenchment to the statements in the evidence set to make some parts of the evidence more robust and resistant to change. To this end we can for example take

$$d(\vec{q}, \vec{p}) := \sqrt{d_i(q_i - p_i)^2}$$

and define $\vec{mcc}(B)$, as the n-vectors

$$\vec{q} \in \{< a_1, \ldots, a_n > \mid \text{there is a probability function } W \text{ on } S\mathcal{L} \text{ with } W(\phi_j) = a_j\}$$

[1] that is there is no probability function that can simultaneously assign these values to the sentences in $\phi_1, \ldots, \phi_{n+1}$.

for which $d(\vec{q}, \vec{p})$ is minimal. And as before let the revision of B by $w(\phi_{n+1}) = p_{n+1}$ be

$$B' = \{w(\phi_1) = q_1, \ldots, w(\phi_n) = q_n, w(\phi_{in+1}) = q_{n+1}\}$$

where

$$\vec{q} = \vec{mcc}(B \cup \{w(\phi_{n+1}) = p_{n+1}\}).$$

One can achieve the same results by taking a more detailed approach using some notion of ordinal ranking. To see this take the language $\mathcal{L}^{(k)}$ to have the same relation symbols as $\mathcal{L}$, say $R_1, \ldots, R_t$ but with the domain restricted to $\{a_1, \ldots, a_k\}$. If k is the largest such that a_k appears in ϕ_i, $i = 1, \ldots, n+1$, then the ϕ_i can be viewed as sentences in the propositional language with propositional variables

$$R_i(a_{j_1}, \ldots, a_{j_{s_i}})$$

with $1 \leq i \leq t$, $i_1, \ldots, i_{s_i} \in \{a_1, \ldots, a_k\}$ and s_i being the arity of R_i. Then the atoms of this language are the sentences of the form

$$\bigwedge_{\substack{j_1,\ldots,j_{s_i} \leq k \\ R\ s_i-ary \\ R_i \in R\mathcal{L}, j \in N^+}} \pm R_i(a_{j_1}, ..., a_{j_{s_i}}).$$

and given an ordinal ranking on these atoms in a way that contradictions are given rank 0, and the more plausible atoms get assigned a higher ordinal, one can take the coefficients d_i above as the highest rank such that there is an atom of that rank consistent with ϕ_i. On other contextual consideration one might choose to have the coefficients d_i to represent the reliability of the source or the process from which the information is acquired.

2.3 Probabilistic Entailment

2.3.1 The $^{\eta}\triangleright_{\zeta}$ Entailment

In this section we will generalise a probabilistic entailment relation introduced by Knight, (Knight 2002), and further developed by Paris (Paris 2004) and Paris, Picado-Muino and Rosefield, (Paris et. al. 2008), and present analogous results to those given by them in the propositional case, for first order languages. Later in this section we shall study a generalisation of this entailment relation following to multiple thresholds as the basis for *reasoning with conflicting evidence* following (Picado-Muino 2008).

As will be clear shortly, the probabilistic entailment we study provides a spectrum of consequence relations, each at a different degree of reliability, which facilitate our goal in deriving meaningful inferences from an inconsistent set. As

we shall see in details, this is in line with our initial thesis to identify an inconsistent theory with an uncertain theory which we shall represent as a probabilistic one. The inferences from such a theory will inevitably be probabilistic and, following (Paris 2004; Paris et. al. 2008), we shall regard the entailment relation as preserving the reliability or "acceptability" of the consequences given that of the premises as opposed to preserving the categorical truth as is the case for the classical consequence relation. The "acceptability" in inferences here, will be represented with a probabilistic threshold which, we shall assume, can be set from the context of the reasoning.

Definition 2.3.1 *(Knight 2002) Let* $\Gamma \subset SL$, $\psi \in SL$ *and* $\eta, \zeta \in [0,1]$.

$$\Gamma^{\eta} \rhd_{\zeta} \psi \iff \textit{for all probability functions } w \textit{ on } \mathcal{L}, \textit{ if } w(\Gamma) \geq \eta \textit{ then } w(\psi) \geq \zeta$$

The idea here is that as long as one is in the position to assign to each of the sentences in Γ a probability of at least η, one is also in the position to assign a probability of at least ζ to the sentence ψ. The intuition for defining such a probabilistic entailment is more evident when $\eta = \zeta$ are interpreted as the thresholds for acceptance. In this situation the entailment relation $\Gamma^{\eta} \rhd_{\eta} \psi$ can be read as: as long as we are prepared to *accept* all the sentences in Γ we are bound to *accept* ψ. There are situations, however, where the context of reasoning justifies different threshold for the assumptions and conclusion.

An important feature of this entailment relation, relevant to our purpose here, is the observation that for the right value of η this is a para-consistent entailment relation. To see this notice for example that

$$\{\phi, \neg\phi, \psi\}^{1/2} \rhd_{1/2} \neg\psi$$

for ϕ and ψ syntactically disjoint (i.e., when they do not share any relation or constant symbols), since one can find a probability function w for which, $w(\phi) = w(\neg\phi) = 1/2$ and $w(\psi) = 1$ (and thus $w(\neg\psi) = 0$). This does however depend for each Γon the value of η. For $\eta > 1/2$, for example, ${}^{\eta} \rhd_{\zeta}$ will be trivialised on the set $\{\phi, \neg\phi, \psi\}$ for any ζ since there would be no probability function that can assign a probability higher than $1/2$ to all the sentences in this set. To be more precise, the entailment relation ${}^{\eta} \rhd_{\zeta}$ is para-consistent on the set of sentences Γ for all $\eta \leq mc(\Gamma)$. Thus for the rest of this section we shall restrict ourselves to $\eta \in [0, mc(\Gamma)]$ whenever we make a reference to $\Gamma^{\eta} \rhd_{\zeta}$.

Next we shall see some properties of this entailment relation before generalising to the case of multiple thresholds and return to our main goal of providing meaningful logical inference from an inconsistent theory. Many of these properties are generalised from the propositional case, given in (Paris 2004) and (Paris et. al. 2008), immediately and some need modifications to the proof to work for the first order languages. We shall give the proof for the first order case where such modifications are required.

2.3.2 Properties of $^{\eta}\triangleright_{\zeta}$

Proposition 2.3.2 *For any* $\Gamma \subset S\mathcal{L}$ *and* $\psi \in S\mathcal{L}$,
(i) $\Gamma^{\eta} \triangleright_0 \psi$.
(ii) For $\zeta > 0$, $\Gamma^1 \triangleright_{\zeta} \psi \iff \Gamma \models \psi$.
(iii) For $\eta > mc(\Gamma)$, $\Gamma^{\eta} \triangleright_1 \psi$.
(iv) For $\zeta > 0$, $\Gamma^0 \triangleright_{\zeta} \psi \iff \models \psi$.

Proof Parts (i) and (iii) are immediate from the definition. Notice that classical valuations on $\mathcal{L}$ are themselves probability functions. Thus for consistent Γ, $\Gamma^1 \triangleright_{\zeta} \psi$ implies that $v(\psi) \geq \zeta$ for all valuations v for which $v(\Gamma) = 1$. Since $\zeta > 0$ this implies that $v(\psi) = 1$ and thus $\Gamma \models \psi$. If Γ is inconsistent then (ii) follows trivially. Conversely suppose $\Gamma \models \psi$ and $w(\Gamma) = 1$. Let β_i, $1 \leq i \leq m$, enumerate sentences of the form

$$\bigwedge_{i=1}^{n} \phi_i^{\epsilon_i}$$

where $\Gamma = \{\phi_1, \ldots, \phi_n\}$, $\epsilon_i \in \{0, 1\}$ and $\phi_i^1 = \phi_i$ and $\phi_i^0 = \neg\phi_i$. Then for any β_i such that $w(\beta_i) > 0$ we have $\beta_i \models \phi_i$ for all $1 \leq i \leq n$ since otherwise we will have

$$w(\phi_i) = \sum_{\beta_j \models \phi_i} w(\beta_j) < 1.$$

So $\beta_i \models \bigwedge \Gamma$ and since $\bigwedge \Gamma \models \psi$,

$$\zeta \leq 1 = \sum_{\beta_j \models \bigwedge \Gamma} = w(\bigwedge \Gamma) \leq w(\psi)$$

as required. For (iv), if $\not\models \psi$ then there is a valuation v for which $v(\psi) = 0$. Since v is also a probability function and $v(\Gamma) \geq 0$, $\Gamma^0 \triangleright_{\zeta}$ will fail for any $\zeta > 0$. Conversely if $\Gamma^0 \triangleright_{\zeta} \psi$ fails then there si a probability function w for which $w(\psi) < \zeta \leq 1$ and thus $\not\models \psi$. □

Proposition 2.3.3 *Assume that* $\Gamma^{\eta} \triangleright_{\zeta} \psi$. *Then*
(i) If $\tau \geq \eta$ *and* $\nu \leq \zeta$, *then* $\Gamma^{\tau} \triangleright_{\nu} \psi$.
(ii) if $\tau \geq 0$ *and* $\eta + \tau, \zeta + \tau \leq 1$, *then* $\Gamma^{\eta+\tau} \triangleright_{\zeta+\tau} \psi$

We will first prove the following lemma:

Lemma 2.3.4 *Take* $\phi_1, \ldots, \phi_n \in S\mathcal{L}$, *and let* β_i *enumerate the sentences*

$$\bigwedge_{i=1}^{n} \phi_i^{\epsilon_i}$$

as before and let $v(\beta_i)$ *be such that* $\sum_{i=1}^{2^n} v(\beta_i) = 1$. *The there is a probability function,* w *on* $S\mathcal{L}$ *for which*

$$w(\beta_i) = v(\beta_i).$$

Proof It is only enough to define w on $QFS\mathcal{L}$, the quantifier free sentences of $\mathcal{L}$. Choose any probability function u on $S\mathcal{L}$ such that $u(\beta_i) \neq 0$ for $i = 1, \ldots, 2^n$ and for $\psi \in QFS\mathcal{L}$, define

$$w(\psi) = \sum_{i=1}^{2^n} v(\beta_i) u(\psi|\beta_i).$$

□

Proof of Proposition (2.3.3). (i) is immediate from the definition. For (ii) suppose that $\Gamma^{\eta+\tau} \triangleright_{\zeta+\tau} \psi$ failed. Thus there is a probability function w for which $w(\Gamma) \geq \eta + \tau$ but $w(\psi) < \zeta + \tau$. If $w(\psi) < \zeta$ we will have that $\Gamma^{\eta} \triangleright_{\zeta} \psi$ fails. Otherwise let $\gamma \geq 0$ be such that

$$\gamma < \zeta < \gamma + (\zeta + \tau - w(\psi)).$$

Let β_i enumerate all the sentences of the form

$$\bigwedge_{i=1}^{n} \phi_i^{\epsilon_i} \wedge \psi^{\epsilon_{n+1}}.$$

Pick a β_i such that $w(\beta_i) > 0$ and $\beta_i \not\models \psi$ (such a β_i exists otherwise we should have $w(\psi) = 1$ and $\Gamma^{\eta+\tau} \triangleright_{\zeta+\tau} \psi$ will hold). Define

$$v(\beta_k) = \begin{cases} w(\beta_k).(\gamma/w(\psi)) & \text{if } \beta_k \models \psi, \\ w(\beta_k) & \text{if } \beta_k \nVdash \psi, \beta_k \neq \beta_i, \\ w(beta_i) + w(\psi) - \gamma & \text{if } \beta_k = \beta_i \end{cases}$$

so $\sum_{k=1}^{2^{n+1}} v(\beta_k) = 1$. Using Lemma (2.3.4), we can find a probability function w' on $S\mathcal{L}$ such that $w'(\beta_i) = v(\beta_i)$ for $i = 1, \ldots, 2^n$. Then we have:

$$w'(\psi) = \sum_{\beta_i \models \psi} w'(\beta_i) = \sum_{\beta_i \models \psi} w(\beta_i).\gamma/w(\psi) = \gamma$$

and for $\phi \in \Gamma$ we have

$$w(\phi) - w'(\phi) \leq \sum_{\beta_i \models \phi \wedge \psi} w(\beta_i)(1 - \gamma/w(\psi)) \leq w(\psi) - \gamma$$

because for all other $w'(\beta_k) > w(\beta_k)$. So

$$w'(\phi) \geq \eta + \tau - (w(\psi) - \gamma) > \eta.$$

So we have $w'(\phi_i) > \eta$ while $w'(\psi) = \gamma < \zeta$ which contradicts $\Gamma^{\eta} \triangleright_{\zeta} \psi$. □

Proposition 2.3.5 *If* $\lim_{n\to\infty}\eta_n = \eta$ *and* $\lim_{n\to\infty}\zeta_n = \zeta$ *with* η_n *increasing and* $\Gamma^{\eta_n} \rhd_{\zeta_n} \psi$ *for all* n*, then* $\Gamma^{\eta} \rhd_{\zeta} \psi$*.*

Proof See (Picado-Muino 2008)

The next result shows that the entailment relation ${}^{\eta}\rhd_{\zeta}$ does not depend on the choice of language. More precisely, let $\mathcal{L}_1, \mathcal{L}_2$ be finite first order languages and such that $\Gamma \subset S\mathcal{L}_1 \cap S\mathcal{L}_2$ and $\psi \in S\mathcal{L}_1 \cap S\mathcal{L}_2$, then $w_1(\psi) \geq \zeta$ for every probability function w_1 on $S\mathcal{L}$ such that $w_1(\Gamma) \geq \eta$ if and only if $w_2(\psi) \geq \zeta$ for every probability function w_2 on $S\mathcal{L}$ such that $w_2(\Gamma) \geq \eta$.

Proposition 2.3.6 *The relation* ${}^{\eta}\rhd_{\zeta}$ *is language invariant.*

Proof Let $\Gamma \subset S\mathcal{L}$ and $\psi \in S\mathcal{L}$ such that $\Gamma^{\eta} \rhd_{\zeta} \psi$ for the language $\mathcal{L}$, i.e., for every probability function w on $S\mathcal{L}$ if $w(\Gamma) \geq \eta$ then $w(\psi) \geq \zeta$. It is enough to show that if $\mathcal{L}'$ is a language such that $\mathcal{L} \subset \mathcal{L}'$ then for every probability function w' on $S\mathcal{L}'$, if $w'(\Gamma) \geq \eta$ then $w'(\psi) \geq \zeta$ and conversely.

For the forward direction assume that w' is a probability function on $S\mathcal{L}'$ such that $w'(\Gamma) \geq \eta$ but $w'(\psi) < \zeta$. Let w be the restriction of w' to $S\mathcal{L}$. Then w will be a probability function that agrees with w' on Γ and ψ and thus $\Gamma^{\eta} \rhd_{\zeta} \psi$ will fail in the context of the language $\mathcal{L}$. Conversely let w be a probability function on $S\mathcal{L}$ such that $w(\Gamma) \geq \eta$ but $w(\psi) < \zeta$. Let $\Gamma = \{\phi_1, \ldots, \phi_n\}$ and as before let β_i enumerate the sentences of the form

$$\bigwedge_{i=1}^{n} \phi_i^{\epsilon_i} \wedge \psi^{\epsilon_{i+1}}$$

and we have that

$$w(\psi) = \sum_{\beta_i \models \psi} w(\beta_i) < \zeta.$$

Since $\mathcal{L} \subset \mathcal{L}'$, we have $\beta_i \in S\mathcal{L}'$ and since w is a probability function we have that $\sum_{i=1}^{2^{n+1}} w(\beta_i) = 1$. Using lemma 2.3.4, we can find a probability function w' on $S\mathcal{L}'$ with $w'(\beta_i) = w(\beta_i)$. With the notation of Lemma 2.3.4, for $\phi \in \Gamma$,

$$w'(\phi) = \sum_{i=1}^{2^{n+1}} w(\beta_i) u(\phi|\beta_i) = \sum_{\beta_i \models \phi} w(\beta_i) = w(\phi) \geq \eta$$

and

$$w'(\psi) = \sum_{i=1}^{2^{n+1}} w(\beta_i) u(\psi|\beta_i) = \sum_{\beta_i \models \psi} w(\beta_i) = w(\psi) < \zeta.$$

Hence $\Gamma^{\eta} \rhd_{\zeta} \psi$ fails in the context of language $\mathcal{L}'$. □

2.4 Generalising to Multiple Thresholds; $^{\vec{\eta}}\triangleright_{\zeta}$

The intuition behind the probabilistic entailment as studied in the previous sections is to see the relation as extending the classical relation between the *truth* of two sentences to a relation between their reliability. As mentioned before this relation will only make sense if we restrict ourselves to $\eta \in [0, mc(\Gamma)]$ and in particular when dealing with inconsistent information sets, we will be interested in the case where $\eta = mc(\Gamma)$. This intuitively means that we are interested to investigate the probabilistic inferences from a set Γ if we are ready to accept it with the highest reliability consistently possible. That reliability will of course be 1 for a consistent Γ in which case the set of logical consequences of Γ will coincide with its classical consequences.

This approach however might be too coarse a view in many cases. One such case, for example, is when the statements in Γ are accumulated from different sources and their reliability is inevitably bound by the reliability of the corresponding source. Consider a set Γ where some statements in Γ are proved analytically and some are driven from experiments with a certain degree of reliability or error margin. The relation $^{\eta}\triangleright_{\zeta}$, however fails to distinguish between such statements in Γ and the threshold η is assigned to all the sentences in indiscriminatingly. Although this relation provides the means to derive probabilistic inferences from an inconsistent set, as we shall elaborate more in the next section, it fails to limit the effect of inconsistency to the part of the information that is relevant to the inconsistency as we intended. This is because, the threshold η is assigned to the set Γ as a whole and the presence of inconsistencies in Γ will change the maximal consistency for it as a whole. With this idea in mind, one can set out to generalise this entailment relation to a more fine graded relation that allows distinguishing between different parts of the knowledge base. To this end, we will generalise the relation $^{\eta}\triangleright_{\zeta}$ investigated in the previous section. The idea here is that the entailment relation between the set Γ and a sentence ψ is to account not only for the relation between the reliability ψ and that of Γ as a whole but between ψ and the individual sentences in Γ.

Definition 2.4.1 *Let* $\Gamma = \{\phi_1, \ldots, \phi_n\} \subset S\mathcal{L}$, $\psi \in S\mathcal{L}$ *and* $\vec{\eta} \in [0,1]^n, \zeta \in [0,1]$. *Define*

$$\Gamma^{\vec{\eta}}\triangleright_{\zeta} \psi \iff \textit{for all probability functions } w \textit{ on } \mathcal{L},$$
$$\textit{if } w(\phi_i) \geq_{<} \vec{\eta} >_i \textit{ for } i = 1, \ldots, n \textit{ then } w(\psi) \geq \zeta.$$

As mentioned before this gives more suitable grounds for dealing with inconsistent information (in particular when there are reasons to distinguish the sentences in Γ from the reliability point of view, for instance when such sentences are coming from different sources) by allowing to restrict the effect of inconsistencies to only parts of the information set.

2.5 Reasoning with Inconsistent Information

We are now in a position to address the goals towards which we set out. First we wish to be able to make inferences from an inconsistent set while avoiding trivialisation and secondly to limit the effect of inconsistencies to the parts of the reasoning relevant to them. Following our initial approach we interpret the inconsistencies by the uncertainty that they induce in the information and thus essentially deal with uncertain reasoning when trying to reason from inconsistent or conflicting evidence.

Given a set of sentences $\Gamma \subset S\mathcal{L}$, let $\eta = mc(\Gamma)$ and define

$$\Gamma \mid\!\approx_\zeta \psi \iff \Gamma^\eta \rhd_\zeta \psi.$$

Intuitively we have $\Gamma \mid\!\approx_\zeta \psi$ if assuming the highest reliability for the sentences of Γ, ψ will be at least as reliable as ζ. This gives, for each Γ, a spectrum of inference relations $\mid\!\approx_\zeta$ for $\zeta \in [0,1]$ each at a different degree of reliability. Notice that if we denote the set of consequences of Γ at reliability degree ζ by C^ζ_Γ then for $\zeta \leq \delta$ we have

$$C^\delta_\Gamma \subseteq C^\zeta_\Gamma.$$

This does address our first goal to make valid nontrivial inferences from an inconsistent set. To address the second goal we shall move to the fine graded version of the entailment relation; Given a set of sentences $\Gamma \subset S\mathcal{L}$, with $\vec{mcc}(\Gamma) = \vec{\eta}$, define

$$\Gamma \mid\!\approx_\zeta \psi \iff \Gamma^{\vec{\eta}} \rhd_\zeta \psi.$$

Again, we have a spectrum of entailment relations from the set Γ each at a different degree of reliability in $[0,1]$. To see how this allows limiting the effect of inconsistencies consider the following case; Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be disjoint languages with $\mathcal{L} = \mathcal{L}_1 \cup \mathcal{L}_2$ and let $\Gamma_1 \subset S\mathcal{L}_1$ and $\Gamma_2 \subset S\mathcal{L}_2$ so $\Gamma = \Gamma_1 \cup \Gamma_2 \subset S\mathcal{L}$. Let $\Gamma_1 = \{\phi_1, \ldots, \phi_n\}$ be inconsistent with $\vec{mcc}_{\Gamma_1} =< \eta_1, \ldots, \eta_n >$ and assume that $\Gamma_2 = \{\psi_1, \ldots, \psi_m\}$ is consistent and so $\vec{mcc}_{\Gamma_2} =< \delta_1, \ldots, \delta_m >=< 1, \ldots, 1 >$. Then taking

$$\Gamma = \{\phi_1, \ldots, \phi_n, \psi_1, \ldots, \psi_m\}$$

in this fixed order, we have

$$\vec{mcc}(\Gamma) =< \eta_1, \ldots, \eta_n, 1, \ldots, 1 >,$$

and for $\theta \in S\mathcal{L}_2 \subset S\mathcal{L}$ we have

$$\Gamma \mid\!\approx_\zeta \theta \iff \Gamma_2 \models \theta$$

thus reducing the inference on sentences of $\mathcal{L}_2$ where the relevant knowledge is consistent to the classical inference, hence limiting the pathological effect of the

inconsistency only to inferences on sentences $\mathcal{L}_1$ where the knowledge is inconsistent.

Thus reasoning with conflicting evidence in our approach amounts to first identifying the maximal (probabilistic) consistency of the evidence. This will give the reliability of the evidence which in turn identifies the relevant evaluation functions. In the case of classical consequence relation, logical consequences of a set of sentences are those that get value 1 from all relevant evaluation functions. In the same manner the set of logical consequences of a set Γ in our setting are those that receive a probability higher than a certain threshold by all the relevant evaluation functions: the probability functions that satisfy the reliabilities for the evidence given in the $\vec{mcc}(\Gamma)$.

2.6 Conclusion

We started with two goals. First to study consequence relations that allow inference from inconsistent sets and second, to do so in a manner that allows us to limit the effect of inconsistencies to only parts of the reasoning that is relevant to the inconsistency. We studied a process for revising inconsistent information sets to consistent probabilistic ones. Our approach is to change the evaluation inconsistent information by consistently assigning probabilities that are as close as possible to prior evaluations, thus capturing the idea of minimal change revision. Our approach allows for fine grade analysis of the revision process on individual sentences and to allow for the handling of prioritised belief sets.

Next we investigated a probabilistic entailment relation for first order languages that allows us to make logical inference at different degrees of reliabilities. This entailment relation for the right values of the thresholds will yield a para-consistent consequence relation that provides the setting for reasoning with inconsistent information. We derived some basic properties of this relation and studied a generalisation to multiple thresholds which will facilitate our second goal.

Of course our notion of "closeness" when revising the inconsistent belief can be subject to debate. The use of Euclidean distance was motivated by trying to choose the closest values for all sentences simultaneously. It would be interesting to investigate if other notions of "closeness" can improve this approach. Another interesting aspect which we hope to investigate next is to study how to update the weights associated to the informations while dealing with a prioritised belief sets. Given such a set B if one assigns a certain weight to the information $\phi \in B$ and then receives $\neg\phi$ (again with some weight), it seems reasonable to not only revise the valuation of ϕ (and $\neg\phi$) but also the weights that are assigned to these sentences. One would expect such an analysis to depend on how these weights are interpreted above anything else.

2.7 Appendix

Paris, (Paris 2004) and Paris, Picado-Muino and Rosefield, (Paris et. al. 2008) give an analysis of the $^{\eta}\triangleright_{\zeta}$ relation in classical propositional language as well as a complete proof theory. The analysis follows naturally to the first order case and the proof theory can be easily generalised for first order languages. For the sake of completeness, we repeat this analysis here with slight modifications to work in the first order case.

2.7.1 A Classical Analysis of $^{\eta}\triangleright_{\zeta}$

We will now present an analysis of the entailment relation $^{\eta}\triangleright_{\zeta}$ in classical first order logic the intention behind this analysis becomes evident when we discuss its proof theory in the next section. We start with the case where $\eta, \zeta > 0$ are rational and will generalise to the irrational η and ζ after introducing some more technicalities. Thus let $\eta = c/d$ and $\zeta = e/f$ for $c, d, e, f \in \mathbb{N}$ and assume

$$\phi_1, \ldots, \phi_n^{c/d} \triangleright_{e/f} \psi. \tag{2.1}$$

As usual let $\beta_1, \ldots, \beta_m$ enumerate the sentences of the form

$$\bigwedge_{\phi_i}^{\epsilon_i} \wedge \psi^{\epsilon_{n+1}}.$$

Let $\vec{\phi}_i$ eb the m-vector with the jth coordinate 1 if and only if $\beta_j \vDash \phi_i$ and 0 otherwise (notice that if $\beta_j \nvDash \phi_i$ then $\beta_j \vDash \neg\phi_i$) and define $\vec{\psi}$ the same way. Let

$$\mathbb{W}_m = \{< x_1, \ldots, x_m > \mid x_i \geq 0, \sum x_i = 1\}.$$

Notice that $\mathbb{W}_m$ is in one to one correspondence with the probability functions on $S\mathcal{L}$: using Lemma 2.3.4, every $\vec{v} \in \mathbb{W}_m$ can be extended to a probability function w on $S\mathcal{L}$ for which $w(\beta) =< \vec{v} >_i$ and for every probability function w, $< w(\beta_1), \ldots, w(\beta_m) > \in \mathbb{W}_m$ and and we have

$$w(\phi_i) = \sum_{\beta_j \vDash \phi_i} w(\beta_j) = \vec{\phi}_i \cdot < w(\beta_1), \ldots, w(\beta_m) > .$$

With this setting (2.1) will be equivalent to

$$\text{For all } \vec{w} \in \mathbb{W}_m, \text{ if } \vec{\phi}_i \cdot \vec{w} \geq c/d \text{ for } i = 1, \ldots, n, \text{ then } \vec{\psi} \cdot \vec{w} \geq e/f. \tag{2.2}$$

Let $\vec{1}$ be the m-vector with all coordinates 1, and set,

$$\underline{\vec{\phi}_i} = \vec{\phi}_i - (c/d)\vec{1}, \qquad \underline{\vec{\psi}} = \vec{\psi} - (e/f)\vec{1}$$

then (2.2) can be written as

$$\text{For all } \vec{w} \in \mathbb{W}_m, \text{ if } \underline{\vec{\phi}_i} \cdot \vec{w} \geq 0 \text{ for } i = 1, \ldots, n \text{ then } \underline{\vec{\psi}} \cdot \vec{w} \geq 0. \tag{2.3}$$

This means that $\underline{\vec{\psi}}$ is in the cone

$$\left\{ \sum_{i=1}^{n} a_i \underline{\vec{\phi}_i} + \sum_{j=1}^{m} b_j \vec{e}_j | 0 \leq a_i, b_j \in \mathbb{Q} \right\}$$

where $\vec{e}_j$ are the unit m-vectors. This means that for some $0 \leq a_i, b_j \in \mathbb{Q}$,

$$\underline{\vec{\psi}} = \sum_{i=1}^{m} a_i \underline{\phi_i} + \sum_{j=1}^{\vec{m}} b_j \vec{e}_j. \tag{2.4}$$

If we take M to be the product of the denominators of these a_i's, write the a_i's as N_i/M with $M, Ni \in \mathbb{N}$, remove the rightmost expression and multiply both sides by dM, we can rewrite (2.4) as

$$\sum_{i=1}^{n} N_i (df\vec{\phi}_i - cd\vec{1}) \leq M(df\vec{\phi} - de\vec{1}) \tag{2.5}$$

Setting $\neg\vec{\psi} = \vec{1} - \vec{\psi}$, (2.5) will be equivalent to

$$Mdf\neg\vec{\psi} + \sum_{i=1}^{n} dfN_i\vec{\phi}_i \leq [Md(f-e) + cf\sum_{i=1}^{n} m_i]\vec{1} \tag{2.6}$$

Conversely if (2.6) hlds for some $M, N_1, \ldots, N_n \geq 0$ then this process can be reversed to get back (2.2).

Let $\xi_1, \ldots, \xi_N \in \{\phi_1, \ldots, \phi_n\}$ be such that the sentence ϕ_i appears exactly dfN_i many times among $\xi_1, \ldots, \xi_N$ (so $N = df\sum_i N_i$). If $\beta_k \vDash \neg\psi$, by (**??**), the k-th coordinate of ξ_j is non-zero for at most $-deM + cf\sum_i N_i = (cN - d^2eM)/d$ many j. So

$$\bigvee_{\substack{J \subset \{1,\ldots,N\} \\ |J| > (cN - d^2eM)/d}} \bigwedge_{j \in J} \xi_j \vDash \psi \tag{2.7}$$

On the other hand, if $\beta_k \vDash \psi$ then k-th coordinate of ξ_j is non-zero for at most $(cN - d^2M(f-e))/d$ many j. So

$$\bigvee_{\substack{J \subset \{1,\ldots,N\} \\ |J| > (cN - d^2M(f-e))/d}} \bigwedge_{j \in J} \xi_j \vDash \bot. \tag{2.8}$$

Now set,

$$Z = 1 + (cN - d^2eM)/d,$$

$$T = 1 + (cN - d^2M(f-e))/d$$

So

$$Td(f-e) = fcN - edZ + df$$

and we have $T < Z$. From (2.7) and (**??**),

$$\bigvee_{\substack{J\subset\{1,\dots,N\}\\|J|=T}} \bigwedge_{j\in J} \xi_j \vDash \psi, \tag{2.9}$$

$$\bigvee_{\substack{J\subset\{1,\dots,N\}\\|J|=Z}} \bigwedge_{j\in J} \xi_j \vDash \bot \tag{2.10}$$

$$Td(f-e) = fcN - edZ + df \text{and} T < Z. \tag{2.11}$$

Conversely, if for some $\xi_1,\dots,\xi_N$ and some $Z,T\in\mathbb{N}$ (2.9), (2.10) hold and T and Z are related as in (2.11) then for any β_i, if $\beta_i\vDash\neg\psi$ then $\beta_i\vDash\xi_j$ for at most $T-1$ many j. the same way if $\beta_i\vDash\psi$ there are at most $Z-1$ many such j. So

$$\sum_{j=1}^{N}\vec{\xi}_j \le (T-1)\vec{1} + (Z-T)\vec{\psi}. \tag{2.12}$$

Now let $\vec{w}\in\mathbb{W}_m$ and $\vec{\xi}_j\cdot\vec{w}\ge c/d$ for $j=1,\dots,N$. If we multiply both sides of (2.12) with $\vec{w}$ we get

$$(Z-T)\vec{\psi}\cdot\vec{w} \ge (c/d)N - T + 1.$$

But from (2.11),

$$\frac{(c/d)N - t + 1}{Z-T} = e/f$$

so $\vec{\psi}\cdot\vec{w}\ge e/f$. Thus if (2.9), (2.10) and (2.11) hold then

$$\xi_1,\dots,\xi_N^{c/d} \rhd_{e/f} \psi$$

and conversely if

$$\phi_1,\dots,\phi_n^{c/d} \rhd_{e/f} \psi$$

then there is a θ and sentences $\xi_1,\dots,\xi_N\in\Gamma$ (possibly with repeats) such that $\theta\vDash\psi$ and for some $T,Z\in\mathbb{N}$, (2.9), (2.10) and (2.11) hold.

Theorem 2.7.1 *For $\eta,\zeta\in(0,1]$ and $\phi_1,\dots,\phi_n\in S\mathcal{L}$,*

$$\phi_1,\dots,\phi_n{}^{\eta} \rhd_\zeta \psi \iff \exists\xi_1,\dots,\xi_N\in\{\phi_1,\dots,\phi_n\},\ \textit{and } T,Z\in\mathbb{N}\ \textit{with}$$

$$T(1-\zeta) \leq \eta N - \zeta Z + 1, \quad T < Z \text{ and}$$
$$\bigvee_{\substack{J \subset \{1,\ldots,N\} \\ |J|=T}} \bigwedge_{j \in J} \xi_j \models \psi, \quad \bigvee_{\substack{J \subset \{1,\ldots,N\} \\ |J|=Z}} \bigwedge_{j \in J} \xi_j \models \bot$$

Proof The preceding analysis gives the proof for the case of rational η and ζ. To include irrationalη and ζ, first consider the case where η is irrational but $\zeta \in \mathbb{Q}$ and assume that $\Gamma^{\eta} \triangleright_{\zeta} \psi$. Before preceding to show the result for the case where either η or ζ are irrational we need to introduce some notation and technicalities.

Definition 2.7.2 *For* $\Gamma \subset SL, \psi \in SL$ *and* $\eta \in [0,1]$, *let*

$$\zeta^{\psi}_{\Gamma,\eta} = sup\{\zeta \in [0,1] \,|\, \Gamma^{\eta} \triangleright_{\zeta} \psi\}.$$

Using Propositions 2.3.2 and 2.3.5, it is easy to check that this is well defined and that there is a probability function w for which $w(\Gamma) \geq \eta$ and $w(\psi) = \zeta^{\psi}_{\Gamma,\eta}$. We will first show that for all x in some non-empty neighbourhood $(\eta - \epsilon, \eta + \epsilon)$ we have $\zeta^{\psi}_{\Gamma,x} = q_1 x + q_2$ for some $q_1, q_2 \in \mathbb{Q}$. To show this we will first argue that the set of points $(x, \zeta^{\psi}_{\Gamma,x})$ is convex and then we will show that the function $\zeta^{\psi}_{\Gamma,x}$ is continuous on $[0, mc(\Gamma)]$ and so it should be made up of straight lines $y = q_i 1x + q_i 2$ on this interval. By taking ϵ small enough we will end up on a single one of such straight lines in the interval $(\eta - \epsilon, \eta + \epsilon)$.

First notice that by Proposition 2.3.3, if $x_1 \leq x_2$ then $\zeta^{\psi}_{\Gamma,x_1} \leq \zeta^{\psi}_{\Gamma,x_2}$. Thus $\zeta^{\psi}_{\Gamma,x}$ in creasing in x. Second notice that for $\eta_1 < \eta_2 \leq mc(\Gamma)$ and $0 < \delta < 1$ we can find probability functions w_1 and w_2 such that $w_1(\Gamma) \geq \eta_1$ and $w_1(\psi) = \zeta^{\psi}_{\Gamma,\eta_1}$ and similarly $w_2(\Gamma) \geq \eta_2$ and $w_2(\psi) = \zeta^{\psi}_{\Gamma,\eta_2}$. Take $w = \delta w_1 + (1-\delta) w_2$ and we will have

$$w(\phi) = \delta w_1(\phi) + (1-\delta) w_2(\phi) \geq \delta\eta_1 + (1-\delta)\eta_2$$
$$w(\psi) = \delta w_1(\psi) + (1-\delta) w_2(\psi) = \delta\zeta^{\psi}_{\Gamma,\eta_1} + (1-\delta)\zeta^{\psi}_{\Gamma,\eta_2}$$

Thus we have

$$\zeta^{\psi}_{\Gamma,(\delta\eta_1+(1-\delta)\eta_2)} \leq \delta\zeta^{\psi}_{\Gamma,\eta_1} + (1-\delta)\zeta^{\psi}_{\Gamma,\eta_2}$$

This shows that on $[0, mc(\Gamma)]$, $\zeta^{\psi}_{\Gamma,x}$ is both increasing and convex as a function on x. Thus to show its continuity it would be enough to show that $\lim_{x \to mc(\Gamma)} \zeta^{\psi}_{\Gamma,x} = \zeta^{\psi}_{\Gamma,mc(\Gamma)}$. Using Proposition 2.3.5, we have

$$\lim_{x \to mc(\Gamma)} \zeta^{\psi}_{\Gamma,x} \leq \zeta^{\psi}_{\Gamma,mc(\Gamma)}. \tag{2.13}$$

If $\lim_{x\to mc(\Gamma)} \zeta^{\psi}_{\Gamma,x} < \zeta^{\psi}_{\Gamma,mc(\Gamma)}$ then we can take

$$\lim_{x\to mc(\Gamma)} \zeta^{\psi}_{\Gamma,} < t < \zeta^{\psi}_{\Gamma,mc(\Gamma)}$$

we can then find an increasing sequence r_n converging to $mc(\Gamma)$ and probability functions w_n for which $w_n(\Gamma) \geq r_n$ and $w_n(\psi) < t$. For $\Gamma = \{\phi_1, \ldots, \phi_n\}$ and as usual let β_i, $i = 1, \ldots, m$ enumerate the sentences

$$\bigwedge_i^n \phi_i^{\epsilon_i} \wedge \psi^{\epsilon_{n+1}}$$

and consider the vector

$$\vec{w}_j =< w_j(\beta_1), \ldots, w_j(\beta_m) > .$$

Since $w_j(\beta_1)$ is a bounded sequence it has a convergent subsequence, say $w_{1_j}(\beta_1)$, converging to say, $w(\beta_1)$. Let

$$\vec{w}^1_j =< w^1_j(\beta_1), \ldots, w^1_j(\beta_m) >$$

be a subsequence of $\vec{w}_j$ such that w^1_j is a subsequence of w_{1_j} (so $w^1_j(\beta_1)$ converges to $w(\beta_1)$). The same way we have $w^j_1(\beta_2)$ is a bounded sequence and so has a convergent subsequence, say $w_{2_j}(\beta_2)$, converging to say $w(\beta_2)$ and let

$$\vec{w}^2_j =< w^2_j(\beta_1), \ldots, w^2_j(\beta_m) >$$

be a subsequence of $\vec{w}^1_j$ for which $w^2_j(\beta_2)$ is a subsequence of $w_{2_j}(\beta_2)$ and so converges to $w(\beta_2)$. By the same method we will eventually construct a convergent subsequence of $\vec{w}_j$, namely $\vec{w}^m_j$, that converges to

$$\vec{w} =< w(\beta_1), \ldots, w(\beta_m) > .$$

Using Lemma 2.3.4 we can extend this to a probability function w on $S\mathcal{L}$ and for all $\phi \in \Gamma$

$$w(\phi) = \sum_{\beta_k \vDash \phi} w(\beta_k) = \sum_{\beta_k \vDash \phi} \lim_{j\to\infty} w^m_j(\beta_k) = \lim_{j\to\infty} \sum_{\beta_k \vDash \phi} w_j(\beta_k) = \lim_{j\to\infty} w_j(\phi) \geq \lim_{j\to\infty} r_j = r$$

while

$$\begin{aligned} w(\psi) &= \sum_{\beta_k \vDash \psi} w(\beta_k) = \sum_{\beta_k \vDash \psi} \lim_{j\to\infty} w^m_j(\beta_k) \\ &= \lim_{j\to\infty} \sum_{\beta_k \vDash \psi} w_j(\beta_k) = \lim_{j\to\infty} w_j(\psi) < \lim_{j\to\infty} t = t < \zeta^{\psi}_{\Gamma,mc(\Gamma)} \end{aligned}$$

which is a contradiction. Thus the strict inequality can hold in (2.13) and we have

$$\lim_{x \to mc(\Gamma)} \zeta^{\psi}_{\Gamma,x} = \zeta^{\psi}_{\Gamma,mc(\Gamma)}$$

as required. It only remains to show that the set of pints $(x, \zeta^{\psi}_{\Gamma,x})$ is convex. Take $\Psi(x,y)$ to be a formula in the language $\mathcal{R} =< \mathbb{R}, +, \leq, 0, 1 >$ such that for $\eta, \zeta \in [0,1]$, $\mathcal{R} \vDash \Psi(\eta, \zeta) \iff \zeta^{\psi}_{\Gamma,\eta} = \zeta$. Then since $\mathcal{R}$ admits quantifier elimination and is an elementary extension of $\mathcal{Q} =< \mathbb{Q}, +, \leq, 0, 1 >$ we can suppose that $\Psi(x,y)$ is of the form

$$\bigvee_{i=1}^{s} \bigwedge_{j=1}^{u_s} (m_{ij} y * n_{ij} x + k_{ij})$$

for some $m_{ij}, n_{ij}, k_{ij} \in \mathbb{Z}$, where $*$ is either $<$ or $\leq$. The set of pairs (x,y) for which $\mathcal{R} \vDash \bigwedge_{j=1}^{u_s} (m_{ij} y * n_{ij} x + k_{ij})$ is convex. Since $\zeta^{\psi}_{\Gamma,x}$ is a continuous and convex function of x it must be a straight line $y = q_i 1 x + qi2$ with coefficients $q_i 1, q_i 2 \in \mathbb{Q}$ with x ranging over some proper interval (which we can take to be closed since $\zeta^{\psi}_{\Gamma,x}$ is continuous).

Returning to our proof of Theorem 2.7.1, take η irrational and ζ rational and assume

$$\phi_1, \ldots, \phi_n \,{}^{\eta} \triangleright_{\zeta} \psi,$$

By the discussion above, $\zeta^{\psi}_{\Gamma,x} = q_1 x + q_2$ for all x in some non-empty interval $(\eta - \epsilon, \eta + \epsilon)$. Since $q_1 \eta + q_2$ is irrational we should have $q_1 \eta + q_2 > \zeta$ (notice that $\zeta^{\psi}_{\Gamma,x} = q_1 x + q_2$ is the maximum on all *zeta* for which $\phi_1, \ldots, \phi_n \,{}^{\eta} \triangleright_{\zeta} \psi$ and the equlity cannot hold) so there are $r_1, r_2 \in \mathbb{Q}$ such that $r_1 < \eta, r_2 > \zeta$ and $q_1 r_1 + q_2 > r_2$. Taking r_1 withing the ϵ of η then $\zeta^{\psi}_{\Gamma,x} > r_2$ so from the first case for rational thresholds we can find $\xi_1, \ldots, \xi_N \in \Gamma$ and Z, T such that $T(1 - r_2) \leq r_1 N - r_2 Z + 1, T < Z$

$$\bigvee_{\substack{J \subset \{1,\ldots,N\} \\ |J| = T}} \bigwedge_{j \in J} \xi_j \vDash \psi, \qquad \bigvee_{\substack{J \subset \{1,\ldots,N\} \\ |J| = Z}} \bigwedge_{j \in J} \xi_j \vDash \bot \tag{2.14}$$

and by taking r_1 and r_2 close enough to η and ζ we will have

$$T(1 - \zeta) \leq \eta N - \zeta Z + 1 \tag{2.15}$$

as required by Theorem 2.7.1. Conversely if we have $\xi_1, \ldots, \xi_N \in \Gamma$ and Z, T satisfying (2.14) and (2.15), there must be $r_1 < \eta$ and $r_2 > \zeta$ such that

$$T(1 - r_2) \leq r_1 N - r_2 Z + 1$$

and by the rational case above we should have $\phi_1, \ldots, \phi_n \,{}^{r_1} \triangleright_{r_2} \psi$ and thus by Proposition 2.3.3 we have

$$\phi_1, \ldots, \phi_n \,{}^{\eta} \triangleright_{\zeta} \psi.$$

The third case where $\eta \in \mathbb{Q}$ and $\zeta \notin \mathbb{Q}$ is proved similarly. For the last case where η, ζ are both irrational assume $\Gamma^{\eta} \rhd_{\zeta} \psi$. First notice that if $\zeta^{\psi}_{\Gamma,x} > \zeta$ then we can take a rational $r_2 > \zeta$ and close enough to ζ and proceed as above so we will assume that $\zeta^{\psi}_{\Gamma,x} = q_1\eta + q_2 = \zeta$. The by the discussion above we have that $\zeta^{\psi}_{\Gamma,x} = q_1\eta + q_2$ in some non-empty interval $(\eta+\epsilon, \eta-\epsilon)$, and we can choose $r_1 \in \mathbb{Q}$ in this interval and set $r_2 = q_1 r_1 + q_2$ and by the first case for rational thresholds we have $\xi_1, \ldots, \xi_N \in \Gamma$ and $Z < T$ with $T(1 - r_2) \leq r_1 N - r_2 Z + 1$. We notice that we should have equality here otherwise we could increase r_2 while keeping r_1 fixed and show that $\zeta^{\psi}_{\Gamma,r_1} > r_2$ which contradicts the choice of r_2. These Z and T work for r_1 arbitrarily close to η (and $r_2 = q_1 r_1 + q_2$) and so by taking the limit one can readily check that the same Z and T will satisfy the required inequality also for η and ζ. In the other direction, suppose we have $\xi_1, \ldots \xi_N$ for which (2.13) hold and Z, T that satisfy the required inequalities. Then for rational r_1 close to η and $r_2 \leq \frac{(r_1 N - T + 1)}{(Z - T)}$ close to ζ these same $\xi_1, \ldots, \xi_N$, Z and T will give $\Gamma^{r_1} \rhd_{r_2} \psi$. Since r_1 and r_2 can be made arbitrarily close to η and ζ respectively we can use Proposition 2.3.5 to get $\Gamma^{\eta} \rhd_{\zeta} \psi$.

Chapter 3

Learning Indicative Conditionals

3.1 Introduction

Indicative conditional statements of the form "if A, then B" constitute a substantial part of the evidence that we obtain. But how should we change our beliefs in the light of such evidence? This question has prompted a large literature. However, in a recent survey, (Douven 2012) concludes that a proper general account of probabilistic belief updating by learning (probabilistic) conditional information is still to be formulated. And indeed, all accounts that have been proposed so far have problems. Here are three of them. (For a much more detailed discussion, see (Douven 2012).)

First and most straightforwardly, one might identify the natural language indicative conditional $\mathrm{A} \rightarrow \mathrm{B}$ with the material conditional $\mathrm{A} \supset \mathrm{B}$, which is equivalent to $\neg\mathrm{A} \vee \mathrm{B}$. In a well-known article, Popper and Miller, (Popper & Miller 1983), challenged this proposal with an argument based on the probability calculus. It goes as follows. Consider two propositions A and B and a prior probability distribution P with $0 < P(\mathrm{A}) < 1$ and $P(\mathrm{B}|\mathrm{A}) < 1$. We now learn the indicative conditional $\mathrm{A} \rightarrow \mathrm{B}$, which we express as the material conditional $\mathrm{A} \supset \mathrm{B}$. To update our beliefs, we use Bayesian Conditionalisation, i.e. we calculate the posterior probability $P^*(\mathrm{A}) := P(\mathrm{A}|\neg\mathrm{A} \vee \mathrm{B})$. Interestingly, it turns out that $P^*(\mathrm{A}) < P(\mathrm{A})$. That is, learning that B follows from A always decreases the probability of A if one uses the material conditional. However, there are examples (such as the Sundowners Example discussed below) where the posterior probability is intuitively judged to be greater than or equal to the

prior probability, which renders this proposal untenable as a general account.

Second, Lewis, (Lewis 1976), proposed an account called *imaging*, which requires a possible worlds semantics with similarity relations holding between worlds. On this account an indicative conditional is true if its consequence holds true in the closest possible world where its antecedent is true. Imaging on ϕ then transfers the probability of every world in which ϕ is false to its closest world where ϕ holds. It turns out, however, that this proposal also fails to do justice to some of our intuitive judgments (cf. (Douven & Dietz 2011)).

Third, one constructs the posterior probability distribution by minimising the Kullback-Leibler divergence between the posterior probability distribution and the prior probability distribution, taking the learned information as a constraint (expressed as a conditional probability statement) on the posterior probability distribution into account. This approach has been challenged with several clever examples. The most famous one is perhaps van Fraassen's, (van Fraassen 1981), Judy Benjamin Problem which aims at showing that the proposed method may lead to wrong results. Other examples that challenge the Kullback-Leibler method can be found in the work of Douven and his co-authors, (Douven 2012), (Douven & Dietz 2011) and (Douven & Romeijn 2012). In this chapter, we revisit four of these examples and show that minimising the Kullback-Leibler divergence leads to intuitively correct results *if the corresponding probabilistic model reflects the causal structure of the scenario in question.*

The remainder of this chapter is organised as follows. Section 3.2 introduces the Kullback-Leibler divergence and applies it to probabilistic belief updating. We then present the four challenging examples. Section 3.3 shows how these challenges can be met if the above-mentioned methodology methodology is properly applied. Section 3.4 shows how the effects of disabling conditions can be properly modelled. Finally, Section 3.5 takes stock and comments on the scope of our proposal.

3.2 The Kullback-Leibler Divergence and Probabilistic Updating

The Kullback-Leibler divergence $D_{KL}(P'\|P)$ measures the expected difference in the informativeness of two probability distributions P' and P from the point of view of P'. Let $S_1, \ldots, S_n$ be the possible values of a random variable S over which probability distributions P' and P are defined. The Kullback-Leibler divergence between P' and P is then given by

$$D_{KL}(P'\|P) := \sum_{i=1}^{n} P'(S_i) \log \frac{P'(S_i)}{P(S_i)}\,. \tag{3.1}$$

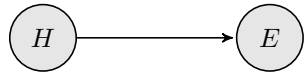


Figure 3.1: The Bayesian Network representation of the relation between H and E.

The Kullback-Leibler divergence is very popular in information theory and has also been used to justify probabilistic updating, (Diaconis & Zabell 1982). Let us show how this works to make ourselves familiar with the Kullback-Leibler divergence and to introduce the methodology that we use in this chapter. To do so, we consider two binary propositional variables.[1] The variable H has two values. H: "The hypothesis holds", and ¬H: "The hypothesis does not hold". The variable E has the values E: "The evidence obtains", and ¬E: "The evidence does not obtain".

We represent the probabilistic dependence between H and E in the Bayesian Network depicted in Figure 1. To complete it, we fix the prior probability of the root node H, i.e.

$$P(\mathrm{H}) = h \tag{3.2}$$

and the conditional probabilities of E, given the values of its parent H:

$$P(\mathrm{E}|\mathrm{H}) = p \quad , \quad P(\mathrm{E}|\neg\mathrm{H}) = q \tag{3.3}$$

The prior probability distribution over H and E is then given by

$$\begin{aligned} P(\mathrm{H},\mathrm{E}) &= h\,p \quad , \quad P(\mathrm{H},\neg\mathrm{E}) = h\,\overline{p} \\ P(\neg\mathrm{H},\mathrm{E}) &= \overline{h}\,q \quad , \quad P(\neg\mathrm{H},\neg\mathrm{E}) = \overline{h}\,\overline{q}\,. \end{aligned} \tag{3.4}$$

Here we have used the convenient shorthand $\overline{x} := 1-x$, which we will use throughout this chapter. We have also used the shorthand notation $P(\mathrm{H},\mathrm{E})$ for $P(\mathrm{H}\wedge\mathrm{E})$ which we will also use below when appropriate.

Next, we learn that the evidence E obtains. This is a constraint on the posterior probability distribution P' which amounts to

$$P'(\mathrm{E}) = 1\,. \tag{3.5}$$

To proceed, we assume that the Bayesian Network depicted in Figure 1 remains unchanged after learning the new information. Hence, the posterior probability

[1] Throughout this chapter we follow the convention, adopted e.g. in Bovens and Hartmann, (Bovens & Hartmann 2003), that propositional variables are printed in (upper case) italic script, and that the instantiations of these variables are printed in (upper case) roman script.

distribution will have the following form:

$$\begin{aligned} P'(\mathrm{H},\mathrm{E}) &= h'\,p' \quad , \quad P'(\mathrm{H},\neg\mathrm{E}) = h'\,\overline{p'} \\ P'(\neg\mathrm{H},\mathrm{E}) &= \overline{h'}\,q' \quad , \quad P'(\neg\mathrm{H},\neg\mathrm{E}) = \overline{h'}\,\overline{q'}\,, \end{aligned} \tag{3.6}$$

where we have replaced all variables by the corresponding primed variables. Eqs. (3.5) and (3.6) then entail that

$$h'\,p' + \overline{h'}\,q' = 1 \tag{3.7}$$

and, taking into account that all four atoms in eqs. (3.6) sum up to 1, that

$$h'\,\overline{p'} = \overline{h'}\,\overline{q'} = 0. \tag{3.8}$$

It is easy to see that eqs. (3.8) only hold for all $h' \in (0,1)$ if $p' = q' = 1$. In this case, eq. (3.7) is automatically fulfilled for all h'. The posterior probability distribution then simplifies as follows:

$$\begin{aligned} P'(\mathrm{H},\mathrm{E}) &= h' \quad , \quad P'(\mathrm{H},\neg\mathrm{E}) = 0 \\ P'(\neg\mathrm{H},\mathrm{E}) &= \overline{h'} \quad , \quad P'(\neg\mathrm{H},\neg\mathrm{E}) = 0 \end{aligned} \tag{3.9}$$

To determine the value of h', we calculate the Kullback-Leibler divergence between P' and P:

$$\begin{aligned} D_{KL}(P'\|P) &:= \sum_{H,E} P'(H,E)\log\frac{P'(H,E)}{P(H,E)} \\ &= h'\log\left(\frac{h'}{h\,p}\right) + \overline{h'}\log\left(\frac{\overline{h'}}{\overline{h}\,q}\right) \\ &= h'\,\log\frac{h'}{h} + \overline{h'}\,\log\frac{\overline{h'}}{\overline{h}} + h'\,\log\frac{q}{p} + \log\frac{1}{q}. \end{aligned} \tag{3.10}$$

We differentiate this expression with respect to h' and obtain after some algebra:

$$\frac{\partial D_{KL}}{\partial h'} = \log\left(\frac{h'}{\overline{h'}}\cdot\frac{\overline{h}}{h}\cdot\frac{q}{p}\right) \tag{3.11}$$

To find the minimum, we set the latter expression equal to zero (i.e. we set the argument of the logarithm equal to 1) and obtain:

$$h' = \frac{h\,p}{h\,p + \overline{h}\,q} \tag{3.12}$$

In more familiar form, this can be written as[2]

$$P'(\mathrm{H}) = P(\mathrm{H}|\mathrm{E}) \equiv P^*(\mathrm{H})\,. \tag{3.13}$$

To complete the proof, we convince ourselves that

$$\frac{\partial^2 D_{KL}}{\partial h'^2} = \frac{1}{h'\,\overline{h'}} > 0 \tag{3.14}$$

for all $h' \in (0,1)$, which shows that we have indeed found the minimum of $D_{KL}(P'||P)$. Hence, Bayes Rule follows from minimising the Kullback-Leibler divergence between the posterior and the prior probability distribution, if one takes the learned information as a constraint on the posterior probability distribution into account.

Let us now explore whether this method can also be used to construct the posterior probability distribution after having learned an indicative conditional (see (Kern-Isberner 2001)). To apply the proposed method, one has to derive a probabilistic statement from the learned conditional.[3] Here we follow (Douven 2012) and others and assume that $P(\mathrm{H} \rightarrow \mathrm{E}) = p$ implies that $P(\mathrm{E}|\mathrm{H}) = p$.[4] In particular, we assume that $\mathrm{H} \rightarrow \mathrm{E}$ implies that $P(\mathrm{E}|\mathrm{H}) = 1$. As in the previous example, the learned conditional is then considered to be a constraint on the posterior probability distribution, which is constructed by minimising the Kullback-Leibler divergence between the posterior probability distribution and the prior probability distribution.

To illustrate our method, let us consider again the Bayesian Network depicted in Figure 1 with the prior probability distribution given in eq. (3.24). Next, we learn that $\mathrm{H} \rightarrow \mathrm{E}$, which implies that

$$P'(\mathrm{E}|\mathrm{H}) := p' = 1. \tag{3.15}$$

[2]In this chapter, P denotes the prior probability distribution, P^* denotes the posterior distribution that follows from Bayesian Conditionalisation, and P' denotes the posterior distribution that follows from minimising the Kullback-Leibler divergence between P' and P satisfying various constraints.

[3]From now on, we drop the adjective "indicative" and the noun "conditional" is always taken to refer to an indicative conditional.

[4]Note that we do not assume that $P(\mathrm{H} \rightarrow \mathrm{E}) = P(\mathrm{E}|\mathrm{H})$. All we need here and indeed throughout the whole chapter is that the learned conditional implies a certain conditional probability constraint on the new probability distribution. And so Lewis' triviality results are of no concern for us.

The Kullback-Leibler divergence between P' and P is then given by

$$\begin{aligned} D_{KL}(P'\|P) &:= \sum_{H,E} P'(H,E) \log \frac{P'(H,E)}{P(H,E)} \\ &= h' \log\left(\frac{h'}{h\,p}\right) + \overline{h'}\left(q' \log\left(\frac{\overline{h'}\,q'}{\overline{h}\,q}\right) + \overline{q'} \log\left(\frac{\overline{h'}\,\overline{q'}}{\overline{h}\,\overline{q}}\right)\right) \qquad (3.16) \\ &= h' \log\frac{h'}{h} + \overline{h'} \log\frac{\overline{h'}}{\overline{h}} + h' \log\frac{1}{p} + \overline{h'}\left(q' \log\frac{q'}{q} + \overline{q'} \log\frac{\overline{q'}}{\overline{q}}\right). \end{aligned}$$

To find the minimum of $D_{KL}(P'\|P)$, we first differentiate this expression with respect to q' and obtain

$$\frac{\partial D_{KL}}{\partial q'} = \overline{h'} \log\left(\frac{q'}{\overline{q'}} \cdot \frac{\overline{q}}{q}\right). \qquad (3.17)$$

Next, we set this expression equal to zero and obtain $q' = q$. With this, we simplify D_{KL} and obtain

$$D_{KL}(P'\|P) = \left(h' \log\frac{h'}{h} + \overline{h'} \log\frac{\overline{h'}}{\overline{h}}\right) + h' \log\frac{1}{p}. \qquad (3.18)$$

Next, we differentiate $D_{KL}(P'\|P)$ with respect to h' and obtain

$$\frac{\partial D_{KL}}{\partial h'} = \log\left(\frac{h'}{\overline{h'}} \cdot \frac{\overline{h}}{h} \cdot \frac{1}{p}\right). \qquad (3.19)$$

Setting this expression equal to zero yields

$$\frac{h'}{\overline{h'}} = p \cdot \frac{h}{\overline{h}}, \qquad (3.20)$$

and hence

$$h' = \frac{hp}{hp + \overline{h}}. \qquad (3.21)$$

Using Lemma 3 from the Appendix, we conclude from eq. (3.20) that $h' < h$, if $0 < p < 1$.[5] This result may sound wrong at first sight. After all, we only learn that H has E as a consequence and nothing else. So why should this prompt us to change our belief in H? And why should the probability of H decrease? Note, however, that H becomes *more informative* after having learned the conditional. If we also learn H, then we can infer with probability 1 that E will obtain as

[5] We skip the proof that the corresponding Hessian is positive definite and that we have therefore found the minimum.

well. It is therefore natural to set the new probability of H to a lower value as more informative hypotheses have a lower probability than less informative hypotheses.[6]

We have seen that minimising the Kulback-Leibler divergence leads to reasonable results for situations involving two propositional variables. But does it also work for more complicated scenarios? Douven and van Fraassen do not think so, and here are four of their alleged counterexamples. Each example starts with a *story* that sets up the scene. Then a conditional is learned which may prompt some previously held beliefs to change.

1. **The Sundowners Example.** Sarah and her sister Marian have arranged to go for sundowners at the Westcliff hotel tomorrow. Sarah feels that there is some chance that it will rain, but thinks they can always enjoy the view from inside. To make sure, Marian consults the staff at the Westcliff hotel and finds out that in the event of rain, the inside area will be occupied by a wedding party. So she tells Sarah: "If it rains tomorrow, then we cannot have sundowners at the Westcliff." Upon learning this conditional, Sarah sets her probability for sundowners *and* rain to 0, but does not change her probability for rain. Thus, in this example, learning the conditional information has the effect of leaving the probability of the antecedent unchanged. This example is from (Douven & Romeijn 2012).

2. **The Ski Trip Example.** Harry sees his friend Sue buying a ski outfit. This surprises him a bit, because he did not know of any plans of hers to go on a ski trip. He knows that she recently had an important exam and thinks it unlikely that she passed it. Then he meets Tom, his best friend and also a friend of Sue's, who is just on his way to Sue to hear whether she passed the exam, and who tells him: "If Sue passed the exam, her father will take her on a ski vacation." Recalling his earlier observation, Harry now comes to find it more likely that Sue passed the exam. So in this example upon learning the conditional information Harry should intuitively increase the probability of the antecedent of the conditional. This example is from (Douven & Dietz 2011).

3. **The Driving Test Example.** Betty knows that Kevin, the son of her neighbours, was to take his driving test yesterday. She has no idea whether or

[6] Note that eq. (3.21) also obtains if one learns the material conditional $\mathrm{H} \supset \mathrm{E} \equiv \neg\mathrm{H} \vee \mathrm{E}$ and uses Bayesian Conditionalisation to update one's beliefs:

$$P^*(\mathrm{H}) = P(\mathrm{H}|\neg\mathrm{H} \vee \mathrm{E}) = \frac{P(\mathrm{H} \wedge (\neg\mathrm{H} \vee \mathrm{E}))}{P(\neg\mathrm{H} \vee \mathrm{E})} = \frac{P(\mathrm{H},\mathrm{E})}{P(\mathrm{H},\mathrm{E}) + P(\neg\mathrm{H})}$$
$$= \frac{hp}{hp + \overline{h}} \equiv h'$$

not Kevin is a good driver; she deems it about as likely as not that Kevin passed the test. Betty notices that her neighbours have started to spade their garden. Then her mother, who is friends with Kevin's parents, calls her and tells her the following: "If Kevin passed the driving test, his parents will throw a garden party." Betty figures that, given the spading that has just begun, it is doubtful (even if not wholly excluded) that a party can be held in the garden of Kevin's parents in the near future. As a result, Betty lowers her degree of belief for Kevin's having passed the driving test and thus decreases the probability of the antecedent of the conditional. This example is from (Douven 2012).

While the first three examples are meant to be challenges to the Kullback-Leibler method, the following problem is an alleged counterexample.

4. **The Judy Benjamin Problem.** A soldier, Judy Benjamin, is dropped with her platoon in a territory that is divided in two halves, Red territory and Blue territory, respectively, with each territory in turn being divided in equal parts, Second Company area and Headquarters Company area, thus forming four quadrants of roughly equal size. Because the platoon was dropped more or less at the centre of the whole territory, Judy Benjamin deems it equally likely that they are in one quadrant as that they are in any of the others. They then receive the following radio message: "I can't be sure where you are. If you are in Red Territory, then the odds are $3:1$ that you are in Second Company area." After this, the radio contact breaks down. Supposing that Judy accepts this message, how should she adjust her degrees of belief?[7]

To address this question, we introduce two binary propositional variables. The variable R has the values R: "Judy lands in Red Territory", and ¬R: "Judy lands in Blue Territory". The variable S has the values S: "Judy lands in Second Company", and ¬S: "Judy lands in Headquarters". The probabilistic dependence between R and S is depicted in the Bayesian Network in Figure 2. To complete it, we fix the prior probability of the root node R, i.e.

$$P(\mathrm{R}) = r \tag{3.22}$$

and the conditional probabilities of S, given the values of its parent R:

$$P(\mathrm{S}|\mathrm{R}) = p \quad , \quad P(\mathrm{S}|\neg\mathrm{R}) = q \tag{3.23}$$

[7] This example is from (van Fraassen 1981).

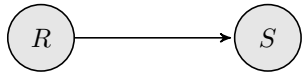


Figure 3.2: The Bayesian Network representation of the relation between R and S.

From the story it is clear that the prior probability distribution over H and E is given by

$$\begin{aligned} P(\mathrm{R},\mathrm{S}) = p\,r = 1/4 \quad &, \quad P(\mathrm{R},\neg\mathrm{S}) = \overline{p}\,r = 1/4 \\ P(\neg\mathrm{R},\mathrm{S}) = q\,\overline{r} = 1/4 \quad &, \quad P(\neg\mathrm{R},\neg\mathrm{S}) = \overline{q}\,\overline{r} = 1/4. \end{aligned} \tag{3.24}$$

Hence,

$$p = q = r = 1/2. \tag{3.25}$$

Next, we learn the conditional "If you are in Red Territory, then the odds are $3:1$ that you are in Second Company area." This is a constraint on the posterior probability distribution P' which amounts to

$$P'(\mathrm{S}|\mathrm{R}) = k\,, \tag{3.26}$$

with $k \in I_{JB} := (0,1) - \{1/2\}$.[8] To proceed, we assume that the Bayesian Network depicted in Figure 2 remains unchanged after learning the new information. Hence, the posterior probability distribution will have the following form:

$$\begin{aligned} P'(\mathrm{R},\mathrm{S}) = r'\,p' \quad &, \quad P'(\mathrm{R},\neg\mathrm{S}) = r'\,\overline{p'} \\ P'(\neg\mathrm{R},\mathrm{S}) = \overline{r'}\,q' \quad &, \quad P'(\neg\mathrm{R},\neg\mathrm{S}) = \overline{r'}\,\overline{q'} \end{aligned} \tag{3.27}$$

As eq. (3.26) implies that $p' = k$, the posterior probability distribution is then given by

$$\begin{aligned} P'(\mathrm{R},\mathrm{S}) = k\,r' \quad &, \quad P'(\mathrm{R},\neg\mathrm{S}) = \overline{k}\,r' \\ P'(\neg\mathrm{R},\mathrm{S}) = \overline{r'}\,q' \quad &, \quad P'(\neg\mathrm{R},\neg\mathrm{S}) = \overline{r'}\,\overline{q'}. \end{aligned} \tag{3.28}$$

We can now calculate the Kullback-Leibler divergence between P' and P and obtain

$$\begin{aligned} D_{KL}(P'\|P) &:= \sum_{R,S} P'(R,S)\log\frac{P'(R,S)}{P(R,S)} \\ &= r'\,\log r' + \overline{r'}\,\log\overline{r'} + r'\,\left(k\,\log k + \overline{k}\,\log\overline{k}\right) \\ &\quad + \overline{r'}\,\left(q'\,\log q' + \overline{q'}\,\log\overline{q'}\right) + \log 4. \end{aligned} \tag{3.29}$$

[8] Note that we consider a more general case here than van Fraassen who focused on $k = 3/4$. We exclude $k = 1/2$ as nothing is learned then.

To find the minimum of $D_{KL}(P'\|P)$, we first calculate its derivative with respect to q',

$$\frac{\partial D_{KL}}{\partial q'} = \overline{r'} \log \frac{q'}{\overline{q'}}, \tag{3.30}$$

and set it to 0. Assuming that $r' \in (0,1)$, we obtain

$$q' = 1/2. \tag{3.31}$$

We now insert eq. (3.31) into eq. (3.29) and differentiate the resulting expression with respect to r'. We then obtain

$$\frac{\partial D_{KL}}{\partial r'} = \log \frac{r'}{\overline{r'}} + \phi(k) \tag{3.32}$$

with

$$\phi(k) := k \log k + \overline{k} \log \overline{k} + \log 2\,. \tag{3.33}$$

To find the minimum, we set the expression in eq. (3.32) equal to 0 and obtain

$$r' = \frac{1}{1 + e^{\phi(k)}}. \tag{3.34}$$

It is easy to see (proof omitted) that $\phi(k) > 0$ for $k \in I_{JB}$. Hence, $r' < 1/2$, i.e. $P'(\mathrm{R}) < P(\mathrm{R})$. However, this is not intuitive, as van Fraassen (1981) has argued: the probability of R should not change after learning the conditional.

In response to van Fraassen, Douven and Romeijn (Douven & Romeijn 2012), have proposed a Bayesian solution of the Judy Benjamin Problem which uses Jeffrey Conditionalisation to model the learning of the uncertain conditional. However, their solution fails to give an adequate account of a number of examples where the probability of the antecedent is intuitively judged to change (as in the previous two examples).

3.3 Meeting the Challenges

To meet the four challenges presented in the previous section, we propose the following methodology. First, we identify all relevant variables of the problem at hand and the causal relations that hold between them. Second, we represent the causal structure by a Bayesian Network and fix the prior probability distribution P that is associated with that network. Third, we express the learned conditional as a constraint on the posterior probability distribution P' and assume that the causal structure is the same before and after learning the conditional. Fourth, we minimise the Kullback-Leibler divergence $D_{KL}(P'\|P)$ between the posterior distribution P' and the prior distribution P to obtain the posterior probability distribution P'. Fifth, we check whether the result complies with our intuitions.

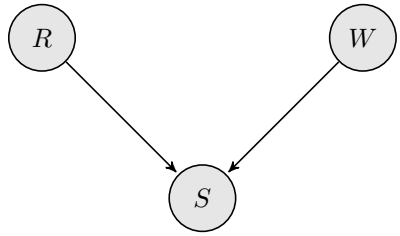


Figure 3.3: The Bayesian Network for the Sundowners Example.

3.3.1 The Sundowners Example

We introduce three binary propositional variables. The variable R has values R: "It will rain tomorrow", and ¬R: "It will not rain tomorrow". The variable W has the values W: "There is a wedding party", and ¬W: "There is no wedding party". Finally, the variable S has the values S: "Sarah and Marian have sundowners", and ¬S: "Sarah and Marian do not have sundowners".

Before we proceed, let us show that using the material conditional and Bayesian Conditionalisation leads to an intuitively wrong result. To do so, remember that Marian tells Sarah "[i]f it rains tomorrow, then we cannot have sundowners at the Westcliff." We formalise this as $\mathrm{R} \supset \neg\mathrm{S}$ which is equivalent to $\neg\mathrm{R} \vee \neg\mathrm{S}$. Using Bayesian Conditionalisation, we then obtain for the posterior probability of R

$$\begin{aligned} P^*(\mathrm{R}) &= P(\mathrm{R}|\neg\mathrm{R} \vee \neg\mathrm{S}) = \frac{P(\mathrm{R} \wedge (\neg\mathrm{R} \vee \neg\mathrm{S}))}{P(\neg\mathrm{R} \vee \neg\mathrm{S})} = \frac{P(\mathrm{R} \wedge \neg\mathrm{S})}{P(\neg\mathrm{R} \vee \neg\mathrm{S})} \\ &= \frac{P(\mathrm{R}) - P(\mathrm{R},\mathrm{S})}{1 - P(\mathrm{R},\mathrm{S})}. \end{aligned} \tag{3.35}$$

Note that the story suggests that $0 < P(\mathrm{R}), P(\mathrm{R},\mathrm{S}) < 1$. Hence, we conclude from eq. (3.35) that $P^*(\mathrm{R}) < P(\mathrm{R})$, which conflicts with our intuitive judgment that the probability of rain should remain unchanged.

Le us now show how our suggested methodology deals with the case. The story suggests a number of dependencies and independencies between the various variables. The Bayesian Network in Figure 3 represents the probabilistic dependencies and independencies between these variables. It also properly represents their causal relations.

To complete the Bayesian Network, we have to fix the prior probability of the root nodes and the conditional probabilities of all other nodes, given the values

of their parents. We set

$$P(\mathrm{R}) = r \quad , \quad P(\mathrm{W}) = w \tag{3.36}$$

and

$$\begin{aligned} P(\mathrm{S}|\mathrm{R},\mathrm{W}) &= \alpha \quad , \quad P(\mathrm{S}|\mathrm{R},\neg\mathrm{W}) = \beta \\ P(\mathrm{S}|\neg\mathrm{R},\mathrm{W}) &= \gamma \quad , \quad P(\mathrm{S}|\neg\mathrm{R},\neg\mathrm{W}) = \delta\,. \end{aligned}$$

If, as we assume, R and W are the only causes of S, then the story suggests that

$$P(\mathrm{S}|\mathrm{R},\mathrm{W}) = \alpha = 0\,. \tag{3.37}$$

All other conditional probabilities (i.e. β, γ and δ) are in the open interval $(0, 1)$. With this, the prior probability distribution over the variables R, S and W has the following form:

$$\begin{aligned} P(\mathrm{R},\mathrm{S},\mathrm{W}) &= 0 \quad , \quad P(\mathrm{R},\neg\mathrm{S},\mathrm{W}) = r\,w \\ P(\mathrm{R},\mathrm{S},\neg\mathrm{W}) &= r\,\overline{w}\,\beta \quad , \quad P(\mathrm{R},\neg\mathrm{S},\neg\mathrm{W}) = r\,\overline{w}\,\overline{\beta} \\ P(\neg\mathrm{R},\mathrm{S},\mathrm{W}) &= \overline{r}\,w\,\gamma \quad , \quad P(\neg\mathrm{R},\neg\mathrm{S},\mathrm{W}) = \overline{r}\,w\,\overline{\gamma} \\ P(\neg\mathrm{R},\mathrm{S},\neg\mathrm{W}) &= \overline{r}\,\overline{w}\,\delta \quad , \quad P(\neg\mathrm{R},\neg\mathrm{S},\neg\mathrm{W}) = \overline{r}\,\overline{w}\,\overline{\delta} \end{aligned} \tag{3.38}$$

Let us now consider the posterior probability distribution P', which is defined over the same Bayesian Network as before. As $\alpha = 0$ (eq. (3.37)), we conclude that

$$\alpha' = 0. \tag{3.39}$$

Another constraint on the posterior probability distribution is the learned conditional "If it rains tomorrow, then we cannot have sundowners at the Westcliff", which implies that

$$P'(\mathrm{S}|\mathrm{R}) = 0 \tag{3.40}$$

and hence $P'(\mathrm{R},\mathrm{S}) = 0$. Using eq. (3.39), this amounts to

$$\overline{w'}\,\beta' = 0\,, \tag{3.41}$$

where we have taken into account that $r' > 0$ as there is no reason (before and after learning the conditional) to assume that the probability of rain is zero. To satisfy eq. (3.41), we are then left with two possibilities: (i) $w' = 1$ and (ii) $\beta' = 0$. It is clear from the story that $\beta = P(\mathrm{S}|\mathrm{R},\neg\mathrm{W}) > 0$: If there is no wedding party, then Sarah and Marian can enjoy their sundowners inside if it rains. This also holds after learning the conditional, hence $\beta' > 0$. Eq. (3.40) then implies that

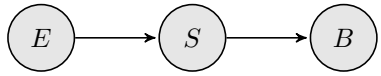


Figure 3.4: The Bayesian Network for the Ski Trip Example.

$w' = 1$. Sarah is now certain that there will be a wedding party. The posterior probability distribution therefore simplifies to

$$\begin{aligned} P'(\mathrm{R},\mathrm{S},\mathrm{W}) &= 0 \quad , \quad P'(\mathrm{R},\neg\mathrm{S},\mathrm{W}) = r' \\ P'(\mathrm{R},\mathrm{S},\neg\mathrm{W}) &= 0 \quad , \quad P'(\mathrm{R},\neg\mathrm{S},\neg\mathrm{W}) = 0 \\ P'(\neg\mathrm{R},\mathrm{S},\mathrm{W}) &= \overline{r'}\,\gamma' \quad , \quad P'(\neg\mathrm{R},\neg\mathrm{S},\mathrm{W}) = \overline{r'}\,\overline{\gamma'} \\ P(\neg\mathrm{R},\mathrm{S},\neg\mathrm{W}) &= 0 \quad , \quad P(\neg\mathrm{R},\neg\mathrm{S},\neg\mathrm{W}) = 0. \end{aligned} \tag{3.42}$$

We can now show the following theorem (proof in the Appendix).

Theorem 3.3.1 *Consider the Bayesian Network depicted in Figure 3 with the prior probability distribution P from eqs. (3.38). We furthermore assume that (i) the posterior probability distribution P' is defined over the same Bayesian Network, (ii) the learned conditional is modelled as a constraint (eq. (3.40)) on P', and (iii) P' minimises the Kullback-Leibler divergence to P. Then $P'(\mathrm{R}) = P(\mathrm{R})$.*

We conclude that the proposed method yields the intuitively correct result in this case.

3.3.2 The Ski Trip Example

Let us first examine the situation before we learn anything. To do so, we introduce the following binary propositional variables. The variable E has the values E: "Sue passed the exam", and ¬E: "Sue did not pass the exam". The variable S has the values S: "Sue's father invites her for a ski trip", and ¬S: "Sue's father does not invite her for a ski trip". The variable B has the values B: "Sue buys a new ski outfit", and ¬B: "Sue does not buy a new ski outfit". The Bayesian Network in Figure 4 represents the probabilistic dependencies and independencies between these variables. It also properly represents the causal relation between these variables.

To complete the Bayesian Network, we have to fix the prior probability of E, i.e.

$$P(\mathrm{E}) = e, \tag{3.43}$$

and the conditional probabilities

$$\begin{aligned} P(\mathrm{S}|\mathrm{E}) &= p_1 \quad , \quad P(\mathrm{S}|\neg\mathrm{E}) = q_1 \\ P(\mathrm{B}|\mathrm{S}) &= p_2 \quad , \quad P(\mathrm{B}|\neg\mathrm{S}) = q_2 . \end{aligned} \tag{3.44}$$

From the story it is clear that $p_1 > q_1$ and $p_2 > q_2$: It is more likely that Sue's father invites her for a ski trip if she passes the exam than if she does not pass the exam. Similarly, it is more likely that Sue buys a new ski outfit if her father invites her for a ski trip than if he does not.

We can now calculate the prior probability distribution over the variables B, E and S:

$$\begin{aligned} P(\mathrm{B},\mathrm{E},\mathrm{S}) &= e\, p_1\, p_2 \quad , \quad P(\neg\mathrm{B},\mathrm{E},\mathrm{S}) = e\, p_1\, \overline{p_2} \\ P(\mathrm{B},\mathrm{E},\neg\mathrm{S}) &= e\, \overline{p}_1 q_2 \quad , \quad P(\neg\mathrm{B},\mathrm{E},\neg\mathrm{S}) = e\, \overline{p}_1\, \overline{q}_2 \\ P(\mathrm{B},\neg\mathrm{E},\mathrm{S}) &= \overline{e}\, q_1\, p_2 \quad , \quad P(\neg\mathrm{B},\neg\mathrm{E},\mathrm{S}) = \overline{e}\, q_1\, \overline{p_2} \\ P(\mathrm{B},\neg\mathrm{E},\neg\mathrm{S}) &= \overline{e}\, \overline{q}_1 q_2 \quad , \quad P(\neg\mathrm{B},\neg\mathrm{E},\neg\mathrm{S}) = \overline{e}\, \overline{q}_1\, \overline{q}_2 \end{aligned} \tag{3.45}$$

Next we learn two items of information, as a result of which our probability distribution changes from P to P'. First, we learn that B obtains. Assuming that the causal structure depicted in Figure 4 does not change, this means that we learn that

$$P'(\mathrm{B}) = e' \left(p'_1\, p'_2 + \overline{p'}_1 q'_2\right) + \overline{e'} \left(q'_1\, p'_2 + \overline{q'}_1 q'_2\right) = 1 \,, \tag{3.46}$$

where we have replaced all variables by the corresponding primed variables. Second, we learn the conditional "if Sue passes the exam, then her father invites her for a ski trip", which implies that

$$P'(\mathrm{S}|\mathrm{E}) = p'_1 = 1. \tag{3.47}$$

Inserting eq. (3.47) into eq. (3.46), we obtain

$$e'\, p'_2 + \overline{e'} \left(q'_1\, p'_2 + \overline{q'_1}\, q'_2\right) = 1. \tag{3.48}$$

This equation only holds for $e' \in (0,1)$, if

$$p'_2 = 1 \tag{3.49}$$

and if

$$q'_1\, p'_2 + \overline{q'_1}\, q'_2 \equiv q'_1 + \overline{q'_1}\, q'_2 = 1.$$

It has the solutions (i) $q'_1 = 1$ and (ii) $q'_2 = 1$. As solution (i) does not make sense, given the story (why should we now be certain that her father invites her for a ski trip if she does not pass the exam?), we conclude that

$$q'_2 = 1. \tag{3.50}$$

Eqs. (3.49) and (3.50) make sure that $P'(\mathrm{B}) = 1$, whether or not Sue's father invites her for a ski trip. Inserting conditions (3.47), (3.49) and (3.50) into the analogues of eqs. (3.45), we can calculate the posterior probability distribution:

$$\begin{aligned} P'(\mathrm{B},\mathrm{E},\mathrm{S}) &= e' \quad , \quad P'(\neg\mathrm{B},\mathrm{E},\mathrm{S}) = 0 \\ P'(\mathrm{B},\mathrm{E},\neg\mathrm{S}) &= 0 \quad , \quad P'(\neg\mathrm{B},\mathrm{E},\neg\mathrm{S}) = 0 \\ P'(\mathrm{B},\neg\mathrm{E},\mathrm{S}) &= \overline{e'}\, q_1' \quad , \quad P'(\neg\mathrm{B},\neg\mathrm{E},\mathrm{S}) = 0 \\ P'(\mathrm{B},\neg\mathrm{E},\neg\mathrm{S}) &= \overline{e'}\,\overline{q'}_1 \quad , \quad P'(\neg\mathrm{B},\neg\mathrm{E},\neg\mathrm{S}) = 0 \end{aligned} \tag{3.51}$$

We can now show the following theorem (proof in the Appendix).

Theorem 3.3.2 *Consider the Bayesian Network in Figure 4 with the prior probability distribution from eq. (3.45). Let*

$$k_0 := \frac{p_1\, p_2}{q_1\, p_2 + \overline{q_1}\, q_2}.$$

We furthermore assume that (i) the posterior probability distribution P' is defined over the same Bayesian Network, (ii) the learned information is modelled as constraints (eqs. (3.46) and (3.47)) on P', and (iii) P' minimises the Kullback-Leibler divergence to P. Then $P'(\mathrm{E}) > P(\mathrm{E})$, iff $k_0 > 1$.

To proceed, we have to explore whether the condition $k_0 > 1$ holds. From the story we learn that Harry thought that it is unlikely that Sue passed the exam, hence e is small. We also learn from the story that Harry is surprised that Sue bought a ski outfit, hence

$$P(\mathrm{B}) = e\,(p_1\, p_2 + \overline{p_1}\, q_2) + \overline{e}\,(q_1\, p_2 + \overline{q_1}\, q_2) \tag{3.52}$$

is small. And as e is small, we conclude that $q_1\, p_2 + \overline{q_1}\, q_2 := \epsilon$ is small. From the story it is also clear that p_2 is fairly large (≈ 1), because Harry did not know of Sue's plans to go skiing, perhaps he even did not know that she is a skier. And so it is very likely that she has to buy a ski outfit to go on the ski trip. At the same time, q_2 will be very small as there is no reason for Harry to expect Sue to buy such a outfit in this case. Finally, p_1 may not be very large, but the previous considerations suggest that $p_1 \gg \epsilon$. We conclude that

$$k_0 = \frac{p_1}{\epsilon} \cdot p_2 \tag{3.53}$$

will typically be greater than 1. If $k_0 \leq 1$, then the probability of E will not increase after learning the two pieces of information. We conclude that the proposed method yields the intuitively correct result in this case.

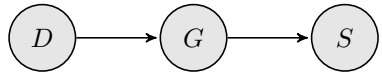


Figure 3.5: The Bayesian Network for the Driving Test Example.

Let us close this subsection with some more general remarks. In Section 3.1 we have seen that one obtains the wrong result if one uses Bayesian Conditionalisation and the material conditional. We showed this in all generality, i.e. it did not matter whether the correct causal structure was taken into account or not. Interestingly, for the Ski Trip Example, it turns out that one gets the right result for the posterior probability of E if one uses the correct causal structure and Bayesian Conditionalisation, i.e. if one updates on B *and* on the material conditional $\mathrm{E} \supset \mathrm{S} \equiv \neg\mathrm{E} \vee \mathrm{S}$. The proof of this result can be found in the Appendix (after the proof of Theorem 2).

3.3.3 The Driving Test Example

Let us first examine the situation before we learn anything. To do so, we introduce the following binary propositional variables. The variable D has the values D: "Kevin passes the driving test", and ¬D: "Kevin does not pass the driving test". The variable G has the values G: "Kevin's parents throw a garden party", and ¬G: "Kevin's parents do not throw a garden party". The variable S has the values S: "Kevin's parents spade their garden", and ¬S: "Kevin's parents do not spade their garden". The Bayesian Network in Figure 5 represents the probabilistic dependencies and independencies between these variables. It also properly represents the causal relation between these variables. Note that the Bayesian Network in Figure 5 has the same structure as the Bayesian Network in Figure 4. Our calculation therefore proceeds as in the previous example.

To complete the Bayesian Network, we have to fix the prior probability of D, i.e.

$$P(\mathrm{D}) = d, \tag{3.54}$$

and the conditional probabilities

$$\begin{aligned} P(G|D) &= p_1 \quad , \quad P(G|\neg D) = q_1 \\ P(S|G) &= p_2 \quad , \quad P(S|\neg G) = q_2. \end{aligned}$$

We can now calculate the prior probability distribution over the variables

D, G and S:

$$\begin{aligned} P(\mathrm{D},\mathrm{G},\mathrm{S}) &= d\,p_1\,p_2 \quad , \quad P(\mathrm{D},\mathrm{G},\neg\mathrm{S}) = d\,p_1\,\overline{p_2} \\ P(\mathrm{D},\neg\mathrm{G},\mathrm{S}) &= d\,\overline{p_1}\,q_2 \quad , \quad P(\mathrm{D},\neg\mathrm{G},\neg\mathrm{S}) = d\,\overline{p_1}\,\overline{q_2} \\ P(\neg\mathrm{D},\mathrm{G},\mathrm{S}) &= \overline{d}\,q_1\,p_2 \quad , \quad P(\neg\mathrm{D},\mathrm{G},\neg\mathrm{S}) = \overline{d}\,q_1\,\overline{p_2} \\ P(\neg\mathrm{D},\neg\mathrm{G},\mathrm{S}) &= \overline{d}\,\overline{q_1}\,q_2 \quad , \quad P(\neg\mathrm{D},\neg\mathrm{G},\neg\mathrm{S}) = \overline{d}\,\overline{q_1}\,\overline{q_2} \end{aligned} \tag{3.55}$$

Next we learn two items of information, as a result of which our probability distribution changes from P to P'. First, we learn that S obtains. Assuming that the causal structure depicted in Figure 5 does not change, this means that we learn that

$$P'(\mathrm{S}) = d'\,(p'_1\,p'_2 + \overline{p'}_1 q'_2) + \overline{d'}\,(q'_1\,p'_2 + \overline{q'}_1 q'_2) = 1\,, \tag{3.56}$$

where we have replaced all variables by the corresponding primed variables. Second, we learn the conditional "if Kevin passed the driving test, his parents will throw a garden party", which implies that

$$P'(\mathrm{G}|\mathrm{D}) = p'_1 = 1. \tag{3.57}$$

Inserting eq. (3.57) into eq. (3.56), we obtain:

$$d'\,p'_2 + \overline{d'}\,\left(q'_1\,p'_2 + \overline{q'_1}\,q'_2\right) = 1 \tag{3.58}$$

This equation only holds for $d' \in (0,1)$, if

$$p'_2 = 1 \tag{3.59}$$

and if

$$q'_1\,p'_2 + \overline{q'_1}\,q'_2 \equiv q'_1 + \overline{q'_1}\,q'_2 = 1.$$

It has the solutions (i) $q'_1 = 1$ and (ii) $q'_2 = 1$. As solution (i) does not make sense intuitively, given the story (unless Kevin's parents would have planned the garden party independently), we conclude that

$$q'_2 = 1. \tag{3.60}$$

Inserting conditions (3.57), (3.59) and (3.60) into the analogues of eqs. (3.55), we can calculate the posterior probability distribution:

$$\begin{aligned} P'(\mathrm{D},\mathrm{G},\mathrm{S}) &= d' \quad , \quad P'(\mathrm{D},\mathrm{G},\neg\mathrm{S}) = 0 \\ P'(\mathrm{D},\neg\mathrm{G},\mathrm{S}) &= 0 \quad , \quad P'(\mathrm{D},\neg\mathrm{G},\neg\mathrm{S}) = 0 \\ P'(\neg\mathrm{D},\mathrm{G},\mathrm{S}) &= \overline{d'}\,q'_1 \quad , \quad P'(\neg\mathrm{D},\mathrm{G},\neg\mathrm{S}) = 0 \\ P'(\neg\mathrm{D},\neg\mathrm{G},\mathrm{S}) &= \overline{d'}\,\overline{q'_1} \quad , \quad P(\neg\mathrm{D},\neg\mathrm{G},\neg\mathrm{S}) = 0 \end{aligned} \tag{3.61}$$

We can now show the following theorem (proof in the Appendix).

Theorem 3.3.3 *Consider the Bayesian Network in Figure 5 with the prior probability distribution from eq. (3.55). Let*

$$k_0 := \frac{p_1\, p_2}{q_1\, p_2 + \overline{q_1}\, q_2}.$$

We furthermore assume that (i) the posterior probability distribution P' is defined over the same Bayesian Network, (ii) the learned information is modelled as constraints (eqs. (3.56) and (3.57)) on P', and (iii) P' minimises the Kullback-Leibler divergence to P. Then $P'(\mathrm{D}) < P(\mathrm{D})$, iff $k_0 < 1$.

Note that it is clear from the story that $q_2 \gg p_2$. Hence,

$$k_0 < \frac{p_1 p_2}{q_1 p_2 + \overline{q_1} p_2} = \frac{p_1 p_2}{p_2} = p_1 < 1. \tag{3.62}$$

We conclude that the posterior probability that Kevin passed the driving test is smaller than the prior probability. The proposed method yields the intuitively correct result in this case.

3.3.4 The Judy Benjamin Problem

In our discussion of this problem in Section 2 we introduced two propositional variables, R and S. Before receiving the radio message, Judy considers the two variables to be probabilistically independent. After receiving the radio message, they became probabilistically dependent. This probabilistic dependence (as well as the probabilistic independence before receiving the message) can be represented in the Bayesian Network in Figure 2. However, as should be clear by now, this Bayesian Network does not reflect the causal relation between the two variables: R does not cause S, and S does not cause R. Hence, there must be another explanation for the probabilistic correlation – a common cause of R and S.

So let us introduce a new binary propositional variables X. Its values could be, for example, that there is wind from a certain direction (X), and that there is no wind from a certain direction (¬X). Note, however, that nothing hinges on the specific values of X. All we need is that Judy believes that a common cause, and not a direct causal relation, explains the learned probabilistic correlation between R and S. The Bayesian Network in Figure 6 represents this situation.

This move suggests the following strategy. The situation after receiving the radio message is represented by the Bayesian Network in Figure 6. The situation before receiving the message can be represented by the Bayesian Network in Figure 2 (with R and S being independent). However, for technical reasons[9]

[9]Calculating the Kullback-Leibler divergence between two probability distributions presupposes that both distributions have the same number of atoms.

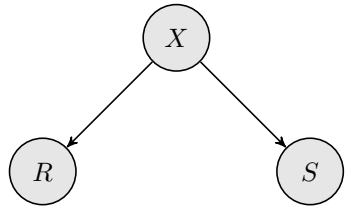


Figure 3.6: The Bayesian Network for the Judy Benjamin Problem.

it is more convenient to also use the Bayesian Network in Figure 6 to represent the situation before receiving the message. All one has to do here is to make sure that R and S are probabilistically independent. One can then determine the posterior probability distribution by minimising the Kullback-Leibler divergence between the posterior and the prior probability distribution and calculate $P(\mathrm{R})$.

So let us proceed and complete the Bayesian Network in Figure 6. First, we have to fix the prior probability of X, i.e.

$$P(\mathrm{X}) = x \in (0,1)\,, \tag{3.63}$$

and the conditional probabilities

$$\begin{aligned} P(\mathrm{R}|\mathrm{X}) &= p_1 \quad , \quad P(\mathrm{R}|\neg\mathrm{X}) = q_1 \\ P(\mathrm{S}|\mathrm{X}) &= p_2 \quad , \quad P(\mathrm{S}|\neg\mathrm{X}) = q_2\,. \end{aligned} \tag{3.64}$$

such that the constraints (see eq. (3.24))

$$\begin{aligned} P(\mathrm{R},\mathrm{S}) &= 1/4 \quad , \quad P(\mathrm{R},\neg\mathrm{S}) = 1/4 \\ P(\neg\mathrm{R},\mathrm{S}) &= 1/4 \quad , \quad P(\neg\mathrm{R},\neg\mathrm{S}) = 1/4 \end{aligned} \tag{3.65}$$

are satisfied. In the Appendix, we show that the following propositions holds.

Proposition 3.3.1 *For the Bayesian Network in Figure 6 and the parameter assignments from eqs. (3.63) and (3.64), the constraints (3.65) imply that (i)* $p_1 = q_1 = 1/2$ *and* $x\,p_2 + \overline{x}\,q_2 = 1/2$ *or (ii)* $p_2 = q_2 = 1/2$ *and* $x\,p_1 + \overline{x}\,q_1 = 1/2$.

To simplify matters, we additionally request that the entropy is maximised, (Williamson 2010).

Proposition 3.3.2 *For the Bayesian Network in Figure 6 and the parameter assignments from eqs. (3.63) and (3.64), the constraints (3.65) imply that setting* $p_1 = q_1 = p_2 = q_2 = x = 1/2$ *maximises the entropy.*

Next Judy learns that "if you are in Red Territory, then the odds are 3 : 1 that you are in Second Company area." This is a constraint on the posterior probability distribution P', which implies that

$$P'(\mathrm{S}|\mathrm{R}) = k\,, \tag{3.66}$$

with $k \in I_{JB}$. In our specific case, we have $k = 3/4$, but (as above) we want to keep things slightly more general by introducing the parameter k. We also notice that the prior probability of X and the likelihoods may change and set

$$P'(\mathrm{X}) = x' \tag{3.67}$$

and

$$\begin{aligned} P'(\mathrm{R}|\mathrm{X}) &=: p_1' \quad , \quad P'(\mathrm{R}|\neg\mathrm{X}) =: q_1' \\ P'(\mathrm{S}|\mathrm{X}) &=: p_2' \quad , \quad P'(\mathrm{S}|\neg\mathrm{X}) =: q_2'\,. \end{aligned} \tag{3.68}$$

Note that x', p_1', q_1', p_2' and q_2' are not independent: They have to satisfy the constraint (3.66).We can now show the following theorem (proof in the Appendix).

Theorem 3.3.4 *Consider the Bayesian Network in Figure 6 with the prior probability distribution P satisfying Proposition 2. We furthermore assume that (i) P' is defined over the same Bayesian Network as P, (ii) the learned information is modelled as a constraint (eqs. (3.66)) on P', and (iii) P' minimises the Kullback-Leibler divergence to P. Then $P'(\mathrm{R}) = P(\mathrm{R})$.*

3.4 Disabling Conditions

The analyses of the examples in the previous section presupposed that all relevant variables can be read off from the story. In particular, we have assumed that there are no interfering causes or disabling conditions that are not mentioned in the story. Our analysis of the Ski Trip Example, for instance, assumed that there is nothing that prevents the father from inviting Sue for a ski trip once he has made the promise. This may not be the case, and we may have beliefs about the presence of a disabling condition. For example, we may consider the possibility that Sue's father changes his mind or that he looses all his money so that he cannot cover the costs of Sue's ski trip anymore. It is also possible that he has an accident or even dies before he can fulfil his promise. In this section, we show how such disabling conditions can be modelled in a straightforward way. To do so, we focus on the Ski Trip Example.

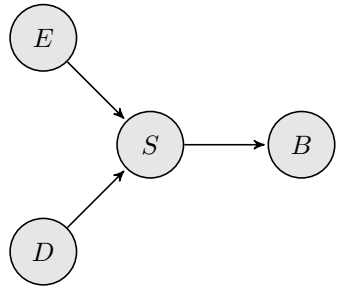


Figure 3.7: The modified Bayesian Network for the Ski Trip Example.

If a disabling condition is present, then the Bayesian Network depicted in Figure 4 has to be modified. In addition to the propositional variables B, E and S from Section 3.2, we add the binary propositional variable D, which is another parent of S. D has the values D: "A disabling condition is present", and ¬D: "No disabling condition is present". The modified Bayesian Network is depicted in Figure 7.

To complete the Bayesian Network, we first fix, as before, the prior probability of E, i.e.

$$P(\mathrm{E}) = e, \tag{3.69}$$

and the conditional probabilities

$$P(\mathrm{B}|\mathrm{S}) = p_2 \quad , \quad P(\mathrm{B}|\neg\mathrm{S}) = q_2. \tag{3.70}$$

Additionally, we fix the prior probability of D, i.e.

$$P(\mathrm{D}) = d, \tag{3.71}$$

and the conditional probabilities

$$\begin{aligned} P(\mathrm{S}|\mathrm{E},\mathrm{D}) &= 0 \quad , \quad P(\mathrm{S}|\mathrm{E},\neg\mathrm{D}) = \beta \\ P(\mathrm{S}|\neg\mathrm{E},\mathrm{D}) &= 0 \quad , \quad P(\mathrm{S}|\neg\mathrm{E},\neg\mathrm{D}) = \delta. \end{aligned} \tag{3.72}$$

This assignment reflects the fact that the presence of a disabling condition prevents the father from inviting Sue for a ski trip. Note that the parameters p_1 and q_1 from eq. (3.44) can be expressed in terms of the new parameters d, β and δ:

$$p_1 = \overline{d}\,\beta \quad , \quad q_1 = \overline{d}\,\delta \tag{3.73}$$

As $p_1 > q_1$, we conclude that $\beta > \delta$.

We can now calculate the prior probability distribution P over the variables B, D, E and S. Here are all non-vanishing atoms:

$$\begin{aligned}
P(\mathrm{E},\mathrm{D},\neg\mathrm{S},\mathrm{B}) &= e\,d\,q_2 \quad , \quad P(\mathrm{E},\mathrm{D},\neg\mathrm{S},\neg\mathrm{B}) = e\,d\,\overline{q_2} \\
P(\mathrm{E},\neg\mathrm{D},\mathrm{S},\mathrm{B}) &= e\,\overline{d}\,\beta\,p_2 \quad , \quad P(\mathrm{E},\neg\mathrm{D},\mathrm{S},\neg\mathrm{B}) = e\,\overline{d}\,\beta\,\overline{p_2} \\
P(\mathrm{E},\neg\mathrm{D},\neg\mathrm{S},\mathrm{B}) &= e\,\overline{d}\,\overline{\beta}\,q_2 \quad , \quad P(\mathrm{E},\neg\mathrm{D},\neg\mathrm{S},\neg\mathrm{B}) = e\,\overline{d}\,\overline{\beta}\,\overline{q_2} \\
P(\neg\mathrm{E},\mathrm{D},\neg\mathrm{S},\mathrm{B}) &= \overline{e}\,d\,q_2 \quad , \quad P(\neg\mathrm{E},\mathrm{D},\neg\mathrm{S},\neg\mathrm{B}) = \overline{e}\,d\,\overline{q_2} \\
P(\neg\mathrm{E},\neg\mathrm{D},\mathrm{S},\mathrm{B}) &= \overline{e}\,\overline{d}\,\delta\,p_2 \quad , \quad P(\neg\mathrm{E},\neg\mathrm{D},\mathrm{S},\neg\mathrm{B}) = \overline{e}\,\overline{d}\,\delta\,\overline{p_2} \\
P(\neg\mathrm{E},\neg\mathrm{D},\neg\mathrm{S},\mathrm{B}) &= \overline{e}\,\overline{d}\,\overline{\delta}\,q_2 \quad , \quad P(\neg\mathrm{E},\neg\mathrm{D},\neg\mathrm{S},\neg\mathrm{B}) = \overline{e}\,\overline{d}\,\overline{\delta}\,\overline{q_2}
\end{aligned} \tag{3.74}$$

Next we learn two items of information, as a result of which our probability distribution changes from P to P'. First, we learn that B obtains. Assuming that the causal structure depicted in Figure 7 does not change, this means that we learn that

$$P'(\mathrm{B}) = p_2'\,\overline{d'}\,(e' + \overline{e'}\,\delta') + q_2'\,(d' + \overline{e'}\,\overline{d'}\,\overline{\delta'}) = 1\,, \tag{3.75}$$

where we have replaced all variables by the corresponding primed variables and assumed that also for the new probability distribution the conditions $P'(\mathrm{S}|\mathrm{E},\mathrm{D}) = P'(\mathrm{S}|\neg\mathrm{E},\mathrm{D}) = 0$ hold. Equation (3.75) only holds for $d', e' \in (0,1)$, if

$$p_2' = q_2' = 1. \tag{3.76}$$

Second, we learn the conditional "if Sue passes the exam, then her father invites her for a ski trip". We interpret this conditional as a *ceteris paribus* claim: If no disabling condition is present, then Sue's father invites her for a ski trip if she passes the exam, i.e.

$$P'(\mathrm{S}|\mathrm{E},\neg\mathrm{D}) = \beta' = 1. \tag{3.77}$$

Inserting conditions (3.76) and (3.77) into the analogues of eqs. (3.74), we can calculate the posterior probability distribution. Again, we only list the non-vanishing atoms:

$$\begin{aligned}
P'(\mathrm{E},\mathrm{D},\neg\mathrm{S},\mathrm{B}) &= e'\,d' \quad , \quad P'(\mathrm{E},\neg\mathrm{D},\mathrm{S},\mathrm{B}) = e'\,\overline{d'} \\
P'(\neg\mathrm{E},\mathrm{D},\neg\mathrm{S},\mathrm{B}) &= \overline{e'}\,d' \quad , \quad P'(\neg\mathrm{E},\neg\mathrm{D},\mathrm{S},\mathrm{B}) = \overline{e'}\,\overline{d'}\,\delta' \\
P'(\neg\mathrm{E},\neg\mathrm{D},\neg\mathrm{S},\mathrm{B}) &= \overline{e'}\,\overline{d'}\,\overline{\delta'}
\end{aligned} \tag{3.78}$$

We can now show the following theorem (proof in the Appendix).

Theorem 3.4.1 *Consider the Bayesian Network in Figure 7 with the prior probability distribution from eq. (3.74). Let*

$$k_d := \frac{p_1\,p_2}{q_1\,p_2 + (\overline{q_1} - d)\,q_2}.$$

We furthermore assume that (i) the posterior probability distribution P' is defined over the same Bayesian Network, (ii) the learned information is modelled as constraints (eqs. (3.76) and (3.77)) on P', and (iii) P' minimises the Kullback-Leibler divergence to P. Then $P'(\mathrm{E}) > P(\mathrm{E})$, iff $k_d > 1$. Moreover, if $k_d > 1$ and $p_2 > q_2$, then $P'(\mathrm{D}) < P(\mathrm{D})$.

Hence, under the conditions discussed above, we expect the probability of E to increase and the probability of D to decrease, which is what we would (or should) expect in this case. We should not be so sure anymore that a disabling condition obtained as the best explanation for the observation that Sue bought a new ski outfit is that she passed the exam (i.e. the probability of E increases) and that her father therefore invited her for a ski trip. Note, finally that in the limit $d \to 0$, Theorem 2 emerges as a special case of Theorem 5.

3.5 Conclusion

We have argued that the Kullback-Leibler divergence minimisation method provides us with an intuitively correct posterior probability distribution if the causal structure of the problem at hand is properly taken into account. We have shown this by giving a detailed account of three challenges and one alleged counterexample that have been discussed in the literature. But does the method also give the right results if more complicated scenarios are considered? We do not see a way how to answer this question in full generality. An answer can probably only be given on a case-by-case basis. We are, however, optimistic that the proposed method will work for more complicated scenarios (which will involve more than three variables) as our examples represent all cases of probabilistic dependencies that can hold between three variables. And so we invite our critics to come up with clever examples where the proposed method fails.

3.6 Appendix

3.6.1 Three Lemmata

The following three lemmata will be useful for the proofs presented in the remainder of this Appendix.

Lemma 1 *Let $f(x) := \log(a\,x), g(x) := x\,\log(a\,x)$ and $h(x) := \overline{x}\,\log(a\,\overline{x})$. Then the first derivatives are: $f'(x) = 1/x, g'(x) = 1 + \log(a\,x)$ and $h'(x) = -1 - \log(a\,\overline{x})$.*

Proof: Trivial.

Lemma 2 *The function* $f(x) := x \log \frac{x}{x'} + \overline{x} \log \frac{\overline{x}}{\overline{x'}}$ *has a minimum at* $x = x'$.

Proof Using Lemma **??**, we obtain

$$f'(x) = \log \left(\frac{x}{\overline{x}} \cdot \frac{\overline{x'}}{x'} \right).$$

Setting this expression equal to zero (i.e. the argument of the logarithm equal to 1), one obtains $x = x'$. As $f''(x) = 1/(x\,\overline{x}) > 0$ for all $x \in (0,1)$, we have indeed found the minimum. □

Lemma 3 *Consider the equation* $x'/\overline{x'} = k \cdot x/\overline{x}$ *with* $k > 0$. *Then (i)* $x' > x$ *iff* $k > 1$, *(ii)* $x' = x$ *iff* $k = 1$ *and (iii)* $x' < x$ *iff* $k < 1$.

Proof This follows from the observation that the function $\varphi(x) := x/\overline{x}$ is strictly monotonically increasing for $x \in (0,1)$. □

3.6.2 Theorem 1

With the prior probability distribution from eq. (3.38) and the posterior probability distribution from eq. (3.42), we obtain for the Kullback-Leibler divergence (3.1) between P' and P:

$$\begin{aligned} D_{KL}(P' \| P) &:= \sum_{R,W,S} P'(R,W,S) \cdot \log \left(\frac{P'(R,W,S)}{P(R,W,S)} \right) \\ &= r' \log \left(\frac{r'}{r\,w} \right) + \overline{r'}\,\gamma' \log \left(\frac{\overline{r'}\,\gamma'}{\overline{r}\,\gamma\,w} \right) + \overline{r'}\,\overline{\gamma'} \log \left(\frac{\overline{r'}\,\overline{\gamma'}}{\overline{r}\,\overline{\gamma}\,w} \right) \\ &= r' \log \frac{r'}{r} + \overline{r'} \log \frac{\overline{r'}}{\overline{r}} + \overline{r'} \left(\gamma' \log \frac{\gamma'}{\gamma} + \overline{\gamma'} \log \frac{\overline{\gamma'}}{\overline{\gamma}} \right) + \log \frac{1}{w} \end{aligned}$$

Next, we differentiate this expression with respect to r' and γ' and obtain

$$\frac{\partial D_{KL}}{\partial r'} = \log \left(\frac{r'}{\overline{r'}} \cdot \frac{\overline{r}}{r} \right) - \left(\gamma' \log \frac{\gamma'}{\gamma} + \overline{\gamma'} \log \frac{\overline{\gamma'}}{\overline{\gamma}} \right) \tag{3.79}$$

$$\frac{\partial D_{KL}}{\partial \gamma'} = \overline{r'} \log \left(\frac{\gamma'}{\overline{\gamma'}} \cdot \frac{\overline{\gamma}}{\gamma} \right). \tag{3.80}$$

Setting the expression in eq. (3.80) equal to zero, we obtain

$$\gamma' = \gamma. \tag{3.81}$$

Substituting this result into eq. (3.79), we obtain

$$\frac{\partial D_{KL}}{\partial r'} = \log\left(\frac{r'}{\overline{r'}} \cdot \frac{\overline{r}}{r}\right). \tag{3.82}$$

Setting the expression in eq. (3.82) equal to zero, we finally obtain $r' = r$. To show that we have indeed found a minimum, we calculate the Hessian matrix of D_{KL} at $(r', \gamma') = (r, \gamma)$ and obtain

$$H(D_{KL})|_{r,\gamma} = \begin{pmatrix} 1/\overline{r} & 0 \\ 0 & r/(\gamma\,\overline{\gamma}) \end{pmatrix}. \tag{3.83}$$

This matrix is positive definite, which completes the proof of Theorem 3.3.1.

3.6.3 Theorem 2

With the prior probability distribution from eq. (3.45) and the posterior probability distribution from eq. (3.51), we obtain for the Kullback-Leibler divergence between P' and P:

$$\begin{aligned} D_{KL}(P'\|P) &:= \sum_{B,E,S} P'(B,E,S) \cdot \log\left(\frac{P'(B,E,S)}{P(B,E,S)}\right) \\ &= e' \log\left(\frac{e'}{e\,p_1\,p_2}\right) + \overline{e'}\,q_1' \log\left(\frac{\overline{e'}\,q_1'}{\overline{e}\,q_1\,p_2}\right) + \overline{e'}\,\overline{q'}_1 \log\left(\frac{\overline{e'}\,\overline{q'}_1}{\overline{e}\,\overline{q_1}\,q_2}\right) \\ &= e' \log\frac{e'}{e} + \overline{e'} \log\frac{\overline{e'}}{\overline{e}} + \overline{e'}\left(q_1' \log\left(\frac{q_1' p_1}{q_1}\right) + \overline{q_1'} \log\left(\frac{\overline{q_1'} p_1 p_2}{\overline{q_1} q_2}\right)\right) \\ &+ \log\frac{1}{p_1 p_2} \end{aligned}$$

Next, we calculate the first derivatives of $D_{KL}(P'\|P)$ with respect to e' and q_1' and obtain after some algebra:

$$\frac{\partial D_{KL}}{\partial e'} = \log\left(\frac{e'}{\overline{e'}} \cdot \frac{\overline{e}}{e} \cdot \frac{1}{k_0}\right) - q_1' \log\left(\frac{q_1'}{\overline{q_1'}} \cdot \frac{\overline{q_1}\,q_2}{q_1\,p_2}\right) \tag{3.84}$$

$$\frac{\partial D_{KL}}{\partial q_1'} = \overline{e'} \log\left(\frac{q_1'}{\overline{q_1'}} \cdot \frac{\overline{q_1}\,q_2}{q_1\,p_2}\right) \tag{3.85}$$

with

$$k_0 := \frac{p_1\,p_2}{q_1\,p_2 + \overline{q_1}\,q_2}. \tag{3.86}$$

To minimize $D_{KL}(P'\|P)$ we first set (3.85) equal to zero (noting that $e' \in (0,1)$) and obtain

$$q_1' = \frac{q_1\,p_2}{q_1\,p_2 + \overline{q_1}\,q_2}. \tag{3.87}$$

With this, we simplify the expression in eq. (3.84) and obtain

$$\frac{\partial D_{KL}}{\partial e'} = \log\left(\frac{e'}{\overline{e'}} \cdot \frac{\overline{e}}{e} \cdot \frac{1}{k_0}\right). \tag{3.88}$$

Setting now also the expression in eq. (3.88) to zero, we obtain

$$\frac{e'}{\overline{e'}} = k_0 \cdot \frac{e}{\overline{e}}. \tag{3.89}$$

Using Lemma 3, we conclude that $e' > e$ iff $k_0 > 1$. This completes the proof of Theorem 3.3.2. (We skip the proof that the corresponding Hessian is positive definite if eqs. (3.87) and (3.89) hold.)

Let us now calculate the posterior probability of E after learning B and the material conditional $\mathrm{E} \supset \mathrm{S} \equiv \neg\mathrm{E} \vee \mathrm{S}$. We obtain

$$\begin{aligned} P^*(\mathrm{E}) &= P(\mathrm{E}|\mathrm{B} \wedge (\neg\mathrm{E} \vee \mathrm{S})) = \frac{P(\mathrm{E} \wedge \mathrm{B} \wedge (\neg\mathrm{E} \vee \mathrm{S}))}{P(\mathrm{B} \wedge (\neg\mathrm{E} \vee \mathrm{S}))} \\ &= \frac{P(\mathrm{B} \wedge \mathrm{E} \wedge \mathrm{S})}{P((\mathrm{B} \wedge \neg\mathrm{E}) \vee (\mathrm{B} \wedge \mathrm{S}))} = \frac{P(\mathrm{B}, \mathrm{E}, \mathrm{S})}{P(\mathrm{B}, \neg\mathrm{E}) + P(\mathrm{B}, \mathrm{S}) - P(\mathrm{B}, \neg\mathrm{E}, \mathrm{S})} \\ &= \frac{P(\mathrm{B}, \mathrm{E}, \mathrm{S})}{P(\mathrm{B}, \neg\mathrm{E}) + P(\mathrm{B}, \mathrm{E}, \mathrm{S})}. \end{aligned} \tag{3.90}$$

With the Bayesian Network depicted in Figure 4 and the prior probability distribution from eq. (3.45), we then obtain

$$P^*(\mathrm{E}) = \frac{e\, p_1\, p_2}{e\, p_1\, p_2 + \overline{e}\,(q_1\, p_2 + \overline{q_1}\, q_2)} = \frac{e\, k_0}{e\, k_0 + \overline{e}} \equiv e' = P'(\mathrm{E}). \tag{3.91}$$

From this equation it is easy to see that $P^*(\mathrm{E}) > P(\mathrm{E})$ iff $k_0 > 1$. Hence, both procedures yield exactly the same result in this case.

3.6.4 Theorem 3

The proof of Theorem 3.3.3 is analogous to the proof of Theorem 3.3.2. With the prior probability distribution from eq. (3.55) and the posterior probability distribution from eq. (3.61), we obtain for the Kullback-Leibler divergence between

P' and P:

$$\begin{aligned} D_{KL}(P'\|P) &:= \sum_{D,G,S} P'(D,G,S)\cdot\log\left(\frac{P'(D,G,S)}{P(D,G,S)}\right) \\ &= d'\,\log\left(\frac{d'}{d\,p_1\,p_2}\right) + \overline{d'}\,q_1'\,\log\left(\frac{\overline{e'}\,q_1'}{\overline{d}\,q_1\,p_2}\right) + \overline{d'}\,\overline{q'}_1\,\log\left(\frac{\overline{d'}\,\overline{q'}_1}{\overline{d}\,\overline{q_1}\,q_2}\right) \\ &= d'\,\log\frac{d'}{d} + \overline{d'}\,\log\frac{\overline{d'}}{\overline{d}} + \overline{d'}\left(q_1'\,\log\left(\frac{q_1'p_1}{q_1}\right) + \overline{q_1'}\,\log\left(\frac{\overline{q_1'}p_1p_2}{\overline{q_1}q_2}\right)\right) \\ &+ \log\frac{1}{p_1p_2} \end{aligned}$$

Next, we calculate the first derivatives of $D_{KL}(P'\|P)$ with respect to d' and q_1' and obtain after some algebra

$$\frac{\partial D_{KL}}{\partial d'} = \log\left(\frac{d'}{\overline{d'}}\cdot\frac{\overline{d}}{d}\cdot\frac{1}{k_0}\right) - q_1'\,\log\left(\frac{q_1'}{\overline{q'}_1}\cdot\frac{\overline{q_1}\,q_2}{q_1\,p_2}\right) \tag{3.92}$$

$$\frac{\partial D_{KL}}{\partial q_1'} = \overline{d'}\,\log\left(\frac{q_1'}{\overline{q'}_1}\cdot\frac{\overline{q_1}\,q_2}{q_1\,p_2}\right), \tag{3.93}$$

with

$$k_0 := \frac{p_1\,p_2}{q_1\,p_2 + \overline{q_1}\,q_2}. \tag{3.94}$$

To minimize $D_{KL}(P'\|P)$ we first set (3.93) equal to zero (noting that $d' \in (0,1)$) and obtain

$$q_1' = \frac{q_1\,p_2}{q_1\,p_2 + \overline{q_1}\,q_2}. \tag{3.95}$$

With this, we simplify the expression in eq. (3.92) and obtain

$$\frac{\partial D_{KL}}{\partial d'} = \log\left(\frac{d'}{\overline{d'}}\cdot\frac{\overline{d}}{d}\cdot\frac{1}{k_0}\right). \tag{3.96}$$

Setting now also the expression in eq. (3.96) to zero, we obtain

$$\frac{d'}{\overline{d'}} = k_0\cdot\frac{d}{\overline{d}}. \tag{3.97}$$

Using Lemma 3, we conclude that $d' < d$ iff $k_0 < 1$. This completes the proof of Theorem 3.3.3. (We skip the proof that the corresponding Hessian is positive definite if eqs. (3.95) and (3.97) hold.)

Again, using the material conditional yields exactly the same result (and the calculation is analogous to the one at the end of the proof of Theorem 3.3.2).

3.6.5 Proposition 1

From the Bayesian Network in Figure 6 and the constraints (3.65), we obtain:

$$\begin{aligned} xp_1p_2 + \overline{x}q_1q_2 &= 1/4 && (3.98)\\ xp_1\overline{p}_2 + \overline{x}q_1\overline{q}_2 &= 1/4 && (3.99)\\ xp_1\overline{p}_2 + \overline{x}q_1\overline{q}_2 &= 1/4 && (3.100)\\ x\overline{p}_1\overline{p}_2 + \overline{x}\overline{q}_1\overline{q}_2 &= 1/4 && (3.101) \end{aligned}$$

We now add eqs. (3.98) and (3.99) as well as eqs. (3.100) and (3.101) and obtain:

$$\begin{aligned} xp_1 + \overline{x}q_1 &= 1/2 && (3.102)\\ x\overline{p}_1 + \overline{x}\overline{q}_1 &= 1/2 && (3.103) \end{aligned}$$

Note that eq. (3.103) follows from eq. (3.102). Similarly, by adding eqs. (3.98) and (3.100) as well as eqs. (3.99) and (3.101), we obtain:

$$xp_2 + \overline{x}q_2 = 1/2 \tag{3.104}$$

We now solve eq. (3.102) for q_1 and eq. (3.104) for q_2 and insert the resulting expressions in eqs. (3.98) to (3.101). In each case, we obtain:

$$(p_1 - 1/2)(p_2 - 1/2) = 0 \tag{3.105}$$

This equation has two solutions, viz. (i) $p_1 = 1/2$ and (ii) $p_2 = 1/2$. Using eqs. (3.102) and (3.104) completes the proof of Proposition 3.3.1.

3.6.6 Proposition 2

We begin with solution (i) from Proposition 3.3.1 and construct the prior probability distribution.

$$\begin{aligned} P(\mathrm{X},\mathrm{R},\mathrm{S}) &= 1/2\,x\,p_2 \quad , \quad P(\mathrm{X},\mathrm{R},\neg\mathrm{S}) = 1/2\,x\,\overline{p}_2\\ P(\mathrm{X},\neg\mathrm{R},\mathrm{S}) &= 1/2\,x\,p_2 \quad , \quad P(\mathrm{X},\neg\mathrm{R},\neg\mathrm{S}) = 1/2\,x\,\overline{p}_2\\ P(\neg\mathrm{X},\mathrm{R},\mathrm{S}) &= 1/2\,\overline{x}\,q_2 \quad , \quad P(\neg\mathrm{X},\mathrm{R},\neg\mathrm{S}) = 1/2\,\overline{x}\,\overline{q}_2\\ P(\neg\mathrm{X},\neg\mathrm{R},\mathrm{S}) &= 1/2\,\overline{x}\,q_2 \quad , \quad P(\neg\mathrm{X},\neg\mathrm{R},\neg\mathrm{S}) = 1/2\,\overline{x}\,\overline{q}_2. \end{aligned} \tag{3.106}$$

With this, we calculate the entropy

$$S = -\sum_{X,R,S} P(X,R,S)\cdot\log P(X,R,S) \tag{3.107}$$

and obtain

$$S = -(x \log x + \overline{x} \log \overline{x}) - x\,(p_2 \log p_2 + \overline{p}_2 \log \overline{p}_2) - \overline{x}\,(q_2 \log q_2 + \overline{q}_2 \log \overline{q}_2) + \log 2. \tag{3.108}$$

We want to maximize S under the constraint

$$x\,p_2 + \overline{x}\,q_2 = 1/2. \tag{3.109}$$

To do so, we use the method of Lagrange multipliers and first calculate the derivates of

$$L = S - \lambda\,(x\,p_2 + \overline{x}\,q_2 - 1/2) \tag{3.110}$$

with respect to p_2 and q_2. We obtain

$$\frac{\partial L}{\partial p_2} = -x\left(\log \frac{p_2}{\overline{p_2}} + \lambda\right) \tag{3.111}$$

$$\frac{\partial L}{\partial q_2} = -\overline{x}\left(\log \frac{q_2}{\overline{q_2}} + \lambda\right). \tag{3.112}$$

Setting these expressions equal to zero and taking into account that $x \in (0,1)$, we obtain

$$p_2 = q_2 = \frac{1}{1 + e^{\lambda}}. \tag{3.113}$$

Inserting this into eq. (3.109), we obtain

$$p_2 = q_2 = 1/2, \tag{3.114}$$

and hence $\lambda = 0$. Inserting all this into eq. (3.110), we obtain

$$L = 2 \log 2 - (x \log x + \overline{x} \log \overline{x}), \tag{3.115}$$

which maximizes at $x = 1/2$ (cf Lemma 2).

The calculation for solution (ii) proceeds analogously for reasons of symmetry. This completes the proof of Proposition 3.3.2.

3.6.7 Theorem 4

Let us first calculate the prior probability distribution over the variables X, R and S with the parameters given in Proposition 3.3.2:

$$\begin{aligned} P(\mathrm{X},\mathrm{R},\mathrm{S}) &= 1/8 \quad , \quad P(\mathrm{X},\mathrm{R},\neg\mathrm{S}) = 1/8 \\ P(\mathrm{X},\neg\mathrm{R},\mathrm{S}) &= 1/8 \quad , \quad P(\mathrm{X},\neg\mathrm{R},\neg\mathrm{S}) = 1/8 \\ P(\neg\mathrm{X},\mathrm{R},\mathrm{S}) &= 1/8 \quad , \quad P(\neg\mathrm{X},\mathrm{R},\neg\mathrm{S}) = 1/8 \\ P(\neg\mathrm{X},\neg\mathrm{R},\mathrm{S}) &= 1/8 \quad , \quad P(\neg\mathrm{X},\neg\mathrm{R},\neg\mathrm{S}) = 1/8 \end{aligned} \tag{3.116}$$

After receiving the message, the probability distribution changes from P to P'. With eqs. (3.67) and (3.68), the posterior probability distribution is given by:

$$\begin{aligned}
P'(\mathrm{X},\mathrm{R},\mathrm{S}) &= x'\,p_1'\,p_2' \quad , \quad P'(\mathrm{X},\mathrm{R},\neg\mathrm{S}) = x'\,p_1'\,\overline{p_2'} \\
P'(\mathrm{X},\neg\mathrm{R},\mathrm{S}) &= x'\,\overline{p_1'}\,p_2' \quad , \quad P'(\mathrm{X},\neg\mathrm{R},\neg\mathrm{S}) = x'\,\overline{p_1'}\,\overline{p_2'} \\
P'(\neg\mathrm{X},\mathrm{R},\mathrm{S}) &= \overline{x'}\,q_1'\,q_2' \quad , \quad P'(\neg\mathrm{X},\mathrm{R},\neg\mathrm{S}) = \overline{x'}\,q_1'\,\overline{q_2'} \\
P'(\neg\mathrm{X},\neg\mathrm{R},\mathrm{S}) &= \overline{x'}\,\overline{q_1'}\,q_2' \quad , \quad P'(\neg\mathrm{X},\neg\mathrm{R},\neg\mathrm{S}) = \overline{x'}\,\overline{q_1'}\,\overline{q_2'}
\end{aligned} \tag{3.117}$$

The parameters x', p_1', q_1', p_2' and q_2' have to be fixed to fit the constraint from eq. (3.66), i.e.

$$\frac{x'\,p_1'\,p_2' + \overline{x}'\,q_1'\,q_2'}{x'\,p_1' + \overline{x}'\,q_1'} = k \tag{3.118}$$

or

$$x'\,p_1'\,(p_2' - k) + \overline{x}'\,q_1'\,(q_2' - k) = 0\,. \tag{3.119}$$

With the prior probability distribution from eq. (3.116) and the posterior probability distribution from eq. (3.117), we obtain for the Kullback-Leibler divergence (3.1) between the two distributions:

$$\begin{aligned}
D_{KL}(P'\|P) &= \sum_{X,R,S} P'(X,R,S)\cdot\log\left(\frac{P'(X,R,S)}{P(X,R,S)}\right) \\
&= \log 8 + x'\log x' + \overline{x'}\log\overline{x'} \\
&\quad + x'\,(p_1'\log p_1' + \overline{p_1'}\log\overline{p_1'}) + \overline{x'}\,(q_1'\log q_1' + \overline{q_1'}\log\overline{q_1'}) \\
&\quad + x'\,(p_2'\log p_2' + \overline{p_2'}\log\overline{p_2'}) + \overline{x'}\,(q_2'\log q_2' + \overline{q_2'}\log\overline{q_2'})
\end{aligned} \tag{3.120}$$

We want to minimize $D_{KL}(P'\|P)$ under the constraint (3.119). To do so, we use the method of Lagrange multipliers and first calculate the derivates of

$$L = D_{KL}(P'\|P) - \lambda\,(x'\,p_1'\,(p_2' - k) + \overline{x}'\,q_1'\,(q_2' - k)) \tag{3.121}$$

with respect to p_1', p_2', q_1' and q_2'. We obtain

$$\frac{\partial L}{\partial p_1'} = x'\left(\log\frac{p_1'}{\overline{p_1'}} - \lambda(p_2' - k)\right) \quad , \quad \frac{\partial L}{\partial p_2'} = x'\left(\log\frac{p_2'}{\overline{p_2'}} - \lambda p_1'\right) \tag{3.122}$$

$$\frac{\partial L}{\partial q_1'} = \overline{x'}\left(\log\frac{q_1'}{\overline{q_1'}} - \lambda(q_2' - k)\right) \quad , \quad \frac{\partial L}{\partial q_2'} = \overline{x'}\left(\log\frac{q_2'}{\overline{q_2'}} - \lambda q_1'\right). \tag{3.123}$$

To find the minimum, we have to set these expressions equal to zero. We recall

that $x' \in (0,1)$ and obtain

$$\log \frac{p_1'}{\overline{p_1'}} = \lambda(p_2' - k) \quad , \quad \log \frac{p_2'}{\overline{p_2'}} = \lambda p_1' \tag{3.124}$$

$$\log \frac{q_1'}{\overline{q_1'}} = \lambda(q_2' - k) \quad , \quad \log \frac{q_2'}{\overline{q_2'}} = \lambda q_1'. \tag{3.125}$$

To proceed, we assume that p_1', p_2', q_1' and q_2' are in $(0,1)$ and note that $\lambda \neq 0$, because $\lambda = 0$ and eqs. (3.124) and (3.125) imply that $p_1 = p_2 = q_1 = q_2 = 1/2$, which contradicts the constraint in eq. (3.119).

We will next solve the second equation in (3.124) for p_1' and insert the result in the first equation. After some algebra, we obtain

$$\log \frac{p_2'}{\overline{p_2'}} = \frac{\lambda}{1 + e^{-\lambda(p_2' - k)}}. \tag{3.126}$$

Proceeding in the same way with the two equations in (3.125), we obtain

$$\log \frac{q_2'}{\overline{q_2'}} = \frac{\lambda}{1 + e^{-\lambda(q_2' - k)}}. \tag{3.127}$$

Comparing eqs. (3.126) and (3.127), we see that p_2' and q_2' satisfy the same equation. We therefore conjecture that $p_2' = q_2'$. From eqs. (3.124) and (3.125), we then obtain

$$p_1' = q_1'. \tag{3.128}$$

Inserting this into eq. (3.119), we obtain

$$p_2' = q_2' = k \tag{3.129}$$

and hence, from eqs. (3.124) and (3.125), that

$$p_1' = q_1' = 1/2. \tag{3.130}$$

From eq. (3.126), we then obtain $\lambda = 2 \log(k/\overline{k})$. Let us now insert eqs. (3.128) and (3.129) into eq. (3.121). The expression L then simplifies to

$$L = \log 4 + x' \log x' + \overline{x'} \log \overline{x'} + k \log k + \overline{k} \log \overline{k}\,, \tag{3.131}$$

from which we infer that the minimum is at

$$x' = 1/2, \tag{3.132}$$

hence, $x' = x$. We have therefore shown that $p_1' = q_1' = 1/2$, $p_2' = q_2' = k$ and $x' = 1/2$ minimises $D_{KL}(P'||P)$. For these parameter values, we obtain that $P'(\mathrm{R}) = x'p_1' + \overline{x'}q_1' = xp_1 + \overline{x}q_1 = P(\mathrm{R})$, which is the desired result.

To complete the proof, we have to show that there is no other set of parameters (p_1', q_1', p_2', q_2' and x') which minimises $D_{KL}(P'\|P)$. Let us first ask whether there is such a parameter set which does not satisfy the condition $p_2' = q_2' = k$. That is, are there parameter sets for which either (i) $p_2' > k$ and $q_2' < k$, or (ii) $p_2' < k$ and $q_2' > k$, or (iii) $p_2' > k$ and $q_2' > k$, or (iv) $p_2' < k$ and $q_2' < k$, or (v) $p_2' = k$ and $q_2' \neq k$, or (vi) $p_2' \neq k$ and $q_2' = k$? Options (iii), (iv) to (vi) drop out as they violate the constraint (3.119), because we assume that $p_1', q_1', p_2', q_2' > 0$. This leaves us with options (i) and (ii). Next, we define

$$\alpha := -\frac{p_1'}{q_1'} \cdot \frac{p_2' - k}{q_2' - k}. \tag{3.133}$$

From options (i) or (ii) we conclude that $\alpha > 0$. The constraint (3.119) then implies that $\overline{x'} = \alpha\, x'$ and therefore

$$x' = \frac{1}{1+\alpha} \quad \text{and} \quad \overline{x'} = \frac{\alpha}{1+\alpha}\,. \tag{3.134}$$

We also introduce the new variables

$$\begin{aligned} \phi_p &:= p_1' \log p_1' + \overline{p_1'} \log \overline{p_1'} + p_2' \log p_2' + \overline{p_2'} \log \overline{p_2'} \\ \phi_q &:= q_1' \log q_1' + \overline{q_1'} \log \overline{q_1'} + q_2' \log q_2' + \overline{q_2'} \log \overline{q_2'} \end{aligned} \tag{3.135}$$

and express D_{KL} in terms of the new variables α, ϕ_p, ϕ_q and the old variable q_2'. As we will see, D_{KL} does not explicitly depend on q_2'. (Note that we could have chosen any one of the four variables p_1', q_1', p_2' and q_2'.)

$$D_{KL} = \log 8 - \log(1+\alpha) + \frac{1}{1+\alpha}\,\left(\alpha \log \alpha + \phi_p + \alpha\, \phi_q\right) \tag{3.136}$$

We differentiate with respect to α and obtain

$$\frac{\partial D_{KL}}{\partial \alpha} = \frac{1}{(1+\alpha)^2} \cdot \left(\log \alpha - \phi_p + \phi_q\right). \tag{3.137}$$

Setting this expression equal to zero yields

$$\alpha = e^{\phi_p - \phi_q}, \tag{3.138}$$

which we insert into eq. (3.136) to obtain

$$D_{KL} = \log 8 + \phi_p - \log(1 + e^{\phi_p - \phi_q}). \tag{3.139}$$

To find the minimum of this expression, we differentiate with respect to ϕ_p, ϕ_q and q_2':

$$\frac{\partial D_{KL}}{\partial \phi_p} = \frac{1}{1+e^{\phi_p-\phi_q}} \tag{3.140}$$

$$\frac{\partial D_{KL}}{\partial \phi_q} = \frac{e^{\phi_p-\phi_q}}{1+e^{\phi_p-\phi_q}} \tag{3.141}$$

$$\frac{\partial D_{KL}}{\partial q_2'} = \frac{\partial D_{KL}}{\partial \phi_q}\cdot\frac{\partial \phi_q}{\partial q_2'} = \frac{e^{\phi_p-\phi_q}}{1+e^{\phi_p-\phi_q}}\cdot\log\frac{q_2'}{\overline{q_2'}} \tag{3.142}$$

To find a minimum, we set these derivatives equal to zero. Note, however, that since $-2\log 2 \le \phi_p, \phi_q < 0$, we have $1/4 < e^{\phi_p-\phi_q} < 4$. Hence, all derivatives are positive and D_{KL} has no minimum if, as we assumed, either (i) $p_2' > k$ and $q_2' < k$ or (ii) $p_2' < k$ and $q_2' > k$ holds. Hence, all parameter sets which minimise D_{KL} satisfy the condition $p_2' = q_2' = k$. Since $\lambda \neq 0$, we infer from the first equations in (3.124) and (3.125) that $p_1' = q_1' = 1/2$ and, following the same reasoning as above, that $x' = 1/2$. This completes the proof of Theorem 3.3.4. (We skip the proof that the corresponding Hessian is positive definite if eqs. (3.129), (3.130) and (3.132) hold.)

In closing, we note that to show that $P'(\mathrm{R}) = P(\mathrm{R})$, it would have been enough to show that $p_1' = q_1'$. In future work we will examine whether this conclusion obtains for the weaker assumptions formulated in Proposition 3.3.1.

3.6.8 Theorem 5

With the prior probability distribution from eq. (3.74) and the posterior probability distribution from eq. (3.78), we obtain for the Kullback-Leibler divergence between the two distributions:

$$\begin{aligned} D_{KL}(P'\|P) &:= \sum_{E,D,S,B} P'(E,D,S,B)\cdot\log\left(\frac{P'(E,D,S,B)}{P(E,D,S,B)}\right) \\ &= \left(e'\log\frac{e'}{e}+\overline{e'}\log\frac{\overline{e'}}{\overline{e}}\right)+\left(d'\log\frac{d'}{d}+\overline{d'}\log\frac{\overline{d'}}{\overline{d}}\right)+d'\log\frac{1}{q_2} \\ &+ e'\,\overline{d'}\log\frac{1}{\beta\,p_2}+\overline{e'}\,\overline{d'}\left(\delta'\log\frac{\delta'}{\delta\,p_2}+\overline{\delta'}\log\frac{\overline{\delta'}}{\overline{\delta}\,q_2}\right) \end{aligned}$$

Next, we calculate the first derivative of $D_{KL}(P'\|P)$ with respect to δ' and obtain

$$\frac{\partial D_{KL}}{\partial \delta'} = \log\left(\frac{\delta'}{\overline{\delta'}}\cdot\frac{\overline{\delta}}{\delta}\cdot\frac{q_2}{p_2}\right). \tag{3.143}$$

Setting this expression equal to zero yields

$$\delta' = \frac{\delta\, p_2}{\delta\, p_2 + \overline{\delta}\, q_2}. \tag{3.144}$$

Note that $\delta' > \delta$ for $p_2 > q_2$. Plugging eq. (3.144) into eq. (3.143), we obtain:

$$\begin{aligned} D_{KL}(P'\|P) \;=\; & \left(e' \log \frac{e'}{e} + \overline{e'} \log \frac{\overline{e'}}{\overline{e}}\right) + \left(d' \log \frac{d'}{d} + \overline{d'} \log \frac{\overline{d'}}{\overline{d}}\right) + d' \log \frac{1}{q_2} \\ & + e'\,\overline{d'} \log \frac{1}{\beta\, p_2} - \overline{e'}\,\overline{d'} \log(\delta\, p_2 + \overline{\delta}\, q_2) \end{aligned}$$

Next, we differentiate this expression with respect to e' and d' and obtain after some algebra and after using eqs. (3.73):

$$\frac{\partial D_{KL}}{\partial e'} \;=\; \log\left(\frac{e'}{\overline{e'}} \cdot \frac{\overline{e}}{e}\right) - \overline{d'} \log k_d \tag{3.145}$$

$$\frac{\partial D_{KL}}{\partial d'} \;=\; \log\left(\frac{d'}{\overline{d'}} \cdot \frac{\overline{\Delta}}{\Delta}\right) + e' \log k_d \tag{3.146}$$

with

$$k_d := \frac{p_1\, p_2}{q_1\, p_2 + (\overline{q_1} - d)\, q_2} \tag{3.147}$$

and

$$\Delta := \frac{d\, q_2}{d\, q_2 + \overline{d}\,(\delta\, p_2 + \overline{\delta}\, q_2)}. \tag{3.148}$$

Using Lemma 3, eqs. (3.145) and (3.146) then entail that $e' > e$ and $d' < \Delta$ iff $k_d > 1$. If $k_d > 1$ and additionally also $p_2 > q_2$, then it also holds that $d' < d$, because $\Delta < d$ if $p_2 > q_2$. This completes the proof of Theorem 3.4.1. (We skip the proof that the corresponding Hessian is positive definite.)

Chapter 4

Voting, Deliberation And Truth

4.1 Introduction

Consider a group aiming to make a collective decision on a binary choice problem. There are countless examples of such scenarios and different methods and procedures have been proposed and practiced for this purpose in large and small scales. The most suitable procedure will no doubt depend on the type of the group, the kind of problem that the group deals with and the purpose of the collective decision. Thus let's consider a group of autonomous, independent decision makers who make non-strategic individual judgments and who wish to arrive at a collective decision that best approximates some objective truth and, without any attempt to give a precise definition or characterisation of democracy, let's assume that the group wishes to make the decision democratically.

There are at least two ways for the group to come to a collective decision: they can each cast a vote and then use some voting rule to aggregate the votes into a collective choice. The majority rule is one example (probably the most widely practiced example) of a democratic voting rule. Alternatively, the group members can deliberate, they present their arguments and reasons to the others, and eventually arrive at a unanimous verdict – a consensus.

Many examples of collective decision making deal with the aggregation of preferences with no external or objective reference for evaluation. In such cases the decision making procedure is hence preferred on the basis of some collectively agreed upon desiderata. There are, on the other hand, instances that the decision makers aim, or are expected to aim, at the best approximation of some fact or

truth. Juries in court are relevant examples of this: They are expected to convict the guilty, and only the guilty. Expert committees (e.g. on environmental issues) are other cases in point. In these cases, the chosen decision making procedure should facilitate this goal.

Hence, there are (at least) two distinct conceptions of how collective decision rules can be evaluated and justified: the *proceduralist conception* and the *epistemic conception*. According to proceduralist view, the merit of a decision making procedure depends only on its procedural characteristics. The work of Kenneth Arrow and his followers is a famous case in point. According to this view the desirability of a collective decision stems from the procedural characteristics of the method itself and from the desiderata satisfied by the procedure, without any concerns of tracking some independent truth. According to the epistemic conception, on the other hand, the procedural characteristics are not enough for legitimising a decision making procedure. Here the main concern is to apply a method that provides reliable and correct outcomes and a decision making procedure is preferred on the basis of its ability to do just this. One widely debated way to put this comparison in perspective in to ask: What should constitute our main criteria for collective decision making? – The *fairness* of the procedure or the *correctness* of the resulting decision?

Deliberative accounts of democratic decision making (*deliberative democracy*) present an immediate advantage over voting from the proceduralist point of view. This advantage is summarised in the formation of a group consensus. It eliminates the necessity for a compromise that is inevitable in voting scenarios. Indeed the prospect of a collective consensus on which all the group members agree upon is an attractive ideal of the proceduralist account. In this sense the deliberative account performs more favourably than voting from the proceduralist perspective.

On the other hand, voting has strong epistemic support. The literature on the epistemic characteristics of majority voting is extensive, much of which is built on the Condorcet Jury Theorem. In its original form the Condorcet Jury Theorem asserts that for a group of *independent* voters with reliabilities above 50% who are faced with a binary choice problem, the probability that the majority vote coincides with the correct choice increases strictly monotonously with the size of the group and approaches 1 asymptotically. This result has since been improved in the work of social choice theorists who have generalised and modified it to relax the assumptions. For example, List & Goodin generalised the theorem to cases with more than two choice options (List & Goodin 2001). It has also been shown that the conclusion of the theorem holds if one requests that the average reliability of all (independent) group members is greater than 0.5, and in recent work Dietrich and Spiekermann proved a modified version of the Condorcet Jury Theorem where they differentiate between individual dependencies and dependencies on a common cause, (Douven 2012). There is, however, as mentioned

above, a procedural disadvantage to majority voting, namely that the process of voting does not affect the epistemic state of the voters. With this consideration the majority voting will always result in a minority that should comply with the vote of the majority as the collective decision while maintaining their original beliefs. As such it will inevitably result in a compromise or possibly in a persisting conflict.

Our goal in this study is to investigate the deliberative account from an epistemic point of view. We ask: is it also epistemically advantageous to deliberate, or is this procedure only preferable from a proceduralist point of view? The question is whether or not the procedural advantages of the deliberation process can be backed with epistemic support to compare with majority voting. The idea of a deliberation process is to give the group members the opportunity to revise their beliefs in the light of the information they receive from their fellow group members, and ideally, come to an agreement on what the collective decision should be. There is an extensive literature on the epistemic analysis of deliberation, developed in the works of scholars such as Joshua Cohen, David Estlund, and Carlos Nino.[1] This literature is concerned with justifying the deliberative account by claiming epistemic advantage for the decisions made through a deliberation by adhering to the qualitative properties of the process in general; properties such as better availability of information, facilitation of the analysis of arguments and reasons resulting in a higher chance of identifying mistakes and errors, reducing the chance of manipulation by controlling the flow of information, etc.[2]

An important aspect, however, that bears on the epistemic analysis of deliberation, is the way that the deliberation process is carried out. There are of course different ways in which the group members can update their belief based the opinions of others and there are also different attributes of decision makers that can be regarded relevant to the deliberation process. To make an investigation of the epistemic behaviour of the deliberation one thus needs to first decide on how to formulate the deliberation procedure in detail.

There are several attempts in the literature for developing a formal account of rational deliberation, including the Lehrer-Wagner model (Lehrer & Wagner 1981), the Hegselmann-Krause model (Hegselmann & Krause 2002), and more recently the Laputa model developed at Lund University. These models focus on different characteristics of the deliberating group members and deal with different contexts of decision making, but only the Laputa model, amongst these, is built on Bayesian foundations. Olsson and co-authors have developed a detailed Bayesian model for the epistemic

[1] See (Cohen 1989a; Cohen 1989b), (Estlund 1993; Estlund 1994; Estlund 1997), (Manin 1987), and (Nino 1996).

[2] See (Bohman & Rehg 1997), (Cohen 1989a), (Dryzek 1990), (Elster 1998), (Fearon 1998), (Marti 2006), and (Nino 1996).

interaction between the group members in which they allow for consideration of any particular network configuration governing the flow of information in the group. In particular group members might only receive messages from some other group members or have different chances of receiving information from one group member as opposed to another. The model also allows for some personal characteristics of the group members outside the social network, viz. the chance that they would engage an enquiry concerning the matter from some outside source (their *activity*) and the chance that such enquiry would provide them with the right answer (their *aptitude*). This model provides an excellent framework for studying epistemic interactions in cases where group members have limitations (imposed by some network configuration) in accessing each others' opinions. Of course the case where group members receive information from all others is a special case of the model. Here we will not give an overview of the intricacies of the Laputa model and its applications and refer the reader to the literature.[3]

In our own model, however, we focus on cases where the group members receive no information from any outside source and the communication between the different group members is open to all. This is in particular the case for jurors in court or experts on expert panels who are solely interested in making the right decision and who do not make any strategic moves. We will indeed see below that our model of deliberation is inspired by a well-known movie that is set behind the closed doors of the Criminal Court Jury Room: *Twelve Angry Men* (1957). We furthermore introduce another characteristic of the group members in addition to their individual reliability to capture their reliability in assessing the (first order) reliabilities of the other group members ("second order reliability"). Investigating the deliberation process on the basis of these characteristics is the main undertaking of the present study.

Our study is concerned with the epistemic characteristics of the deliberation process and its ability to correctly track the truth. The model works in the context of iterated belief revision where the group members update their opinions in each round by considering the opinions presented by their fellow group members in the previous round and share with the group their updated beliefs repeatedly until the group reaches a consensus. An important justification for the iteration of the updating procedure in our model is the assumption that *the group members become increasingly better in assessing the opinions of others in the course of deliberation.* The idea is that during the deliberation the individual group members discuss their arguments and reasons based on which they can make an assessment of how reliable their fellow group members are. This assessment improves as the deliberation proceeds (as they get more and more information about the others) which in turn makes the iteration of the updating procedure meaningful even if

[3]See (3), (57; Olsson 2011), and (58).

the opinions remain unchanged from one round of deliberation to the next.

The details of the relevant attributes of the decision makers in our model and the updating procedure by which the group members take into account each others' opinions will be presented in details shortly. We would, however, like to emphasise the context for which this process is developed. An important contextual assumption of our model is that group members receive no private information or evidence from any outside source. This means, in particular, that all the group members have access to all the information relevant to the decision. Hence the purpose of the deliberation is to allow the group members to come to a shared interpretation of the evidence by learning and weighting the opinions of their fellow group members and to revise their initial belief so the group converges on a collective decision as a single entity. The case of juries in court can be thought of as an archetypal example of the scenarios we have in mind. If different group members have different evidence bases or if new evidence is fed in during the deliberation, then our model does not apply and other deliberation models have to be constructed to study these scenarios.

The remainder of this paper is organised as follows. In Section 4.2 we will introduce our new Bayesian model of deliberation. In Section 4.3 and 4.4 we will present our main results on the emergence of consensus and the truth tracking properties of the process and will give a comparison of this deliberation process with majority voting from an epistemic point of view. Finally, Section 4.5 concludes and highlights some questions for future research.

4.2 A Bayesian Model of Deliberation

Our aim is to capture the scenario in which a group wishes to deliberate on a binary factual question, say the truth or falsity of some hypothesis. The group members start with some subjective belief based on which every one of them casts a yes/no vote for or against the hypothesis. These verdicts are made public, and every group member gets the chance to present her reasons and arguments for her judgment. These reasons and arguments may lead some group members to revise their initial beliefs based on an evaluation of the reasons and arguments presented by the other group members. In the next round of deliberation, every group member casts a yes/no vote for or against the hypothesis on the basis of her revised beliefs. Again, every group member get the chance to present her reasons and arguments, people revise their beliefs, and so on until a consensus is reached. It should be clear by now that the proposed procedure resembles the one shown in the movie *Twelve Angry Men*.

Let us now start formalising things a bit more: we consider a group of n members which we denote by $a_1, \ldots, a_n$, who deliberate on the truth or falsity of some hypothesis. To proceed we introduce a binary propositional variable H

with the values, H: the hypothesis is true, and ¬H: the hypothesis is false. For reasons of symmetry that will become apparent immediately, we assume that the hypothesis is true. The group members express their individual verdicts in terms of a yes/no vote. The votes are represented by binary propositional variables V_i (for $i = 1, \ldots, n$) with the values: V_i: Group member a_i votes that the hypothesis is true, and $\neg\mathrm{V}_i$: Group member a_i votes that the hypothesis is false.

We start by the same assumptions made for majority voting in the Condorcet Jury Theorem. First, we assume that the votes are independent, given the truth or falsity of the hypothesis, i.e.

$$V_i \perp\!\!\!\perp V_1, \ldots, V_{i-1}, V_{i+1}, \ldots, V_n | H \quad \forall i = 1, \ldots, n. \tag{4.1}$$

Second, we assume that each group member a_i is partially reliable with a reliability r_i defined as follows:

$$r_i := P(\mathrm{V_i}|\mathrm{H}) = P(\neg\mathrm{V_i}|\neg\mathrm{H}). \tag{4.2}$$

That is, we focus on the special case where the rate of false positives equals the rate of false negatives. (This assumption can, of course, be easily relaxed.)

Given this setting, majority voting can be studied and it is easy to see that the probability that the probability that the majority makes the right judgment is given by

$$P_V = \sum_{k=\frac{n+1}{2}}^{n} \sum_{\substack{\{a_{j_1},\ldots,a_{j_k}\} \\ \subset\{a_1,\ldots,a_n\}}} \prod_{t\in \{j_1,\ldots,j_k\}} r_t \prod_{t\notin \{j_1,\ldots,j_k\}} (1 - r_t) \ . \tag{4.3}$$

If all group members are equally likely to make the right individual judgment, i.e. if $r_i =: r$ for all $i = 1, \ldots, n$, then the expression in eq. (4.3) simplifies to

$$P_V = \sum_{k=(n+1)/2}^{n} \binom{n}{k} r^k (1 - r)^{n-k} \ . \tag{4.4}$$

With the help of eqs. (4.3) and (4.4), we can explore the truth-tracking properties of the majority voting procedure. According to the well-known Condorcet Jury Theorem, the expressions in eqs. (4.3) and (4.4) strictly monotonically increase with n and converge to 1, for $r > 0.5$ in eq. (4.4) or when the average of the r_i's is greater than 0.5 for the expression in eq. (4.3).

4.2.1 The Deliberation Procedure

Our deliberation model is defined as a process of iterated belief revision and relies on two characteristics of the decision makers: the (first order) reliabilities (r_i) and the second order reliabilities (c_i). The (first order) reliabilities indicate how

competent the group members are in making the right judgment. This is the same reliability that is used in calculating the probability of correct judgment for the majority voting, which will be useful when we will later compare the voting procedure with the deliberation procedure. The second order reliabilities are considered to characterise the group members' competence in assessing the (first order) reliabilities of the other group members. In the process of deliberation each group member assigns reliabilities to her fellow group members and update her opinion based on these reliabilities (and of course the verdicts of the other group members). A high second order reliability for a_j indicates that the estimated reliabilities that a_j assigns to her fellow group members are closer to their objective reliabilities given by the r_i's. In an ideal situation with $c_j = 1$, for example, the reliability assigned by a_j to her fellow group member a_i will equal a_i's objective reliability, i.e. r_i.

We assume that the values of the r_i's and c_i's are independent of each other. Thus we assume that an individual's ability in assessing the reliabilities of other group members does not depend on her ability to asses the truth or falsity of the hypothesis in question.[4] Moreover, we assume that the group members' reliabilities remain fixed during the course of the deliberation. This means, in particular, that we assume that the group members do not acquire any information from sources outside the group.

Our model is formulated in a Bayesian framework and works as follows: First, every group member casts an initial vote, $V_i^{(0)}$ or $\neg V_i^{(0)}$, for or against the hypothesis in question. We introduce parameters $p_i^{(k)}$ and set $p_i^{(k)} = 1$ if $V_i^{(k)}$ and $p_i^{(k)} = -1$ otherwise. These initial votes, for each person, come from an initial probability assignment $P_i^{(0)}(H)$. We assume that group member i will initially vote V_i if $P_i^{(0)}(H) \geq 0.5$ and $\neg V_i$ otherwise. This relates to the reliabilities in an obvious way, that is, the group member with reliability r_i will assign an initial probability greater or equal to 0.5 (and thus vote correctly) with probability r_i. Next, every member a_i estimates the reliability r_j of her fellow group members a_j, viz.

$$r_{ij}^{(0)} := P_i^{(0)}(\mathrm{V_j}|\mathrm{H}) = P_i^{(0)}(\neg \mathrm{V_j}|\neg \mathrm{H}). \tag{4.5}$$

The higher a_i's second order reliability, the better is a_i's assessment of the reliability of a_j, i.e. the closer is $r_{ij}^{(0)}$ to r_j. For $c_i^{(0)} = 0$, a_i randomly assigns some reliability from the uniform distribution over $(0, 1)$ to a_j (for $j = 1, \ldots, n$), and for $c_i^{(0)} = 1$, we obtain that $r_{ij}^{(0)} = r_j$. Using these reliability estimates, each group

[4] Of course this can be debated. One might argue that someone with a high (first order) reliability has a better knowledge of the issue under discussion and therefore also has a better chance of telling more or less reliable fellow group members apart.

member a_i calculates the *likelihood ratios*[5]

$$x_{ij}^{(0)} := \frac{P_i^{(0)}(\mathrm{V_i}|\neg \mathrm{H})}{P_i^{(0)}(\mathrm{V_i}|\mathrm{H})} = \frac{1 - r_{ij}^{(0)}}{r_{ij}^{(0)}}. \tag{4.6}$$

The revision process is carried out on the basis of the votes casted by the other group members and their estimated likelihood ratios:

$$\begin{aligned} P_i^{(1)}(\mathrm{H}) &= P_i^{(0)}(\mathrm{H}|\mathrm{Vote}_1^{(0)}, \ldots, \mathrm{Vote}_{\mathrm{i-1}}^{(0)}, \mathrm{Vote}_{\mathrm{i+1}}^{(0)}, \ldots, \mathrm{Vote}_{\mathrm{n}}^{(0)}) \\ &= \frac{P_i^{(0)}(\mathrm{H})}{P_i^{(0)}(\mathrm{H}) + (1 - P_i^{(0)}(\mathrm{H})) \prod_{k \neq i=1}^{n} \left(x_{ik}^{(0)}\right)^{p_k}} \end{aligned} \tag{4.7}$$

Here $\mathrm{Vote}_{\mathrm{i}}^{(0)} \in \{V_i, \neg V_i\}$. To derive eq. (4.7), we have assumed independence (which is also assumed in the derivation of the Condorcet Jury Theorem, cf. eq. (4.1)).

The group members will then vote again based on their updated probabilities. As before, a group member votes for the hypothesis if her updated probability is greater than or equal to 0.5, otherwise she votes against it. The next round of deliberation will then start with $P_i^{(1)}(\mathrm{H})$ as prior probabilities, and everybody repeats updating her probability assignments as before considering the new votes. And so on until the votes converge.

4.3 Homogeneous Groups

Let G be a homogeneous group of n members, i.e. a group where all group members have the same reliability. This group deliberates on the truth or falsity of the hypothesis H. We assume that each group member has access (through some shared history for example) to each others' reliabilities (corresponding to $c_i = 1, i = 1, \ldots, n$). We furthermore assume that the group members revise their probability assignment for the truth of the hypothesis using the above procedure. Without loss of generality we assume the hypothesis to be true. Then the following theorem holds.

Theorem 4.3.1 *For a homogeneous group G with reliable group members (i.e. for $r > 0.5$), the following three claims hold:*

[5] We follow the convention used in (Bovens & Hartmann 2003). Note that $r_{ij}^{(0)} \geq 1/2$ implies that $0 \leq x_{ij}^{(0)} \leq 1$ and $r_{ij}^{(0)} \leq 1/2$ implies that $x_{ij}^{(0)} \geq 1$.

(i) The probability that the group reaches a consensus in finitely many steps increases with the size of the group and approaches 1 *as the size of the group increases.*

(ii) If the majority of the group members vote correctly in the first round, the subjective beliefs will stabilise on the truth in finitely many steps, i.e. after finitely many steps, each group member assigns subjective probability 1 *to the truth of the hypothesis after which the deliberation process will no more change the probability assignments.*

(iii) If the majority of the group members vote incorrectly in the first round, the subjective beliefs will stabilise on the wrong belief in finitely many steps, i.e. after finitely many steps, each group member assigns subjective probability 0 *to the truth of the hypothesis after which the deliberation process will no more change the probability assignments.*

Proof See Appendix A1.

For a homogeneous group G with unreliable members, i.e. when all group members have a reliability $r < 0.5$, the situation is more complicated and the emergence of a consensus depends strongly on the size of the group and the initial probabilities. To see this notice that for $r < 0.5$ we will have $x > 1$ and thus $x^{\sum_{j=1}^{n} p_j^{(0)}} < 1$ if and only if $\sum_{j=1}^{n} p_j^{(0)} < 0$, i.e. if the majority of the group members vote incorrectly in the first round. Using the same argument as in the Condorcet Jury Theorem the chance that the majority of the group members (with reliability less than 0.5) will vote incorrectly increases with the size of the group and approaches 1. Thus using the argument in the proof of Theorem 4.3.1 if the majority of the group members start with initial subjective probabilities of less than 0.5 for H and hence vote incorrectly in the first round, the probability assignments will increase in the next round and this continues until at some point, say at round t, the majority assigns a probability greater than 0.5 for H and thus votes correctly. After this stage the process will reverse and the probabilities will start to decrease since $\sum_{j=1}^{n} p_j^{(t)} > 0$ and thus $x^{\sum_{j=1}^{n} p_j^{(t)}} > 1$. If the size of the group, the likelihoods and the initial probabilities are such that at some round $s-1$ the majority assign probabilities less than 0.5 (and thus vote incorrectly) but the probabilities increase in such a way that in round s all the probability assignments are above 0.5 then the group reaches a consensus at this round s. On the other hand if the probability assignments increase until at some round $s-1$ the majority *but not all group members* assign a probability above 0.5 (so the probabilities decrease in the next round) and in round s all probabilities decrease to less than 0.5 then the group will again reach a consensus but this time on the wrong answer. Otherwise the group can oscillate (not necessarily in consecutive

rounds) between the case where a majority vote correctly and the case where the majority vote incorrectly. In any case, the subjective beliefs of the group members will not stabilise for unreliable groups.

Theorem 4.3.2 *For a homogeneous group G with unreliable group members (i.e. for $r < 0.5$), the subjective beliefs of the group members will not stabilise even if the group reaches a consensus.*

Proof See Appendix A2.

Notice that in the proof of Theorem 4.3.1, the actual value of x is not relevant. All that matters is whether $x > 1$ or $x < 1$. This allows for an immediate generalisation of these results.

Corollary 4.3.1 *For a homogeneous group G with first order reliability r, let the second order reliabilities c_i for $i = 1, \ldots, n$ be less than 1 (so the group members won't have access to each others' actual reliabilities) but high enough so that the group members can correctly assess whether or not the other group members are reliable, that is let c_i be high enough so that $r_{ij} > 0.5$ if and only if $r_j > 0.5$ for $j = 1, \ldots, n$. Then the results in Theorems 4.3.1 and 4.3.2 still hold.*

The situation in Theorems 4.3.1 and 4.3.2 is highly idealised as we assume that the second order reliability is 1, which means that the group members have access to each others' objective reliabilities. In such a context it will be hard to justify the iteration of the deliberation process after the second round. Assuming that group members are able to weight each others opinion by the actual objective reliabilities there is no room for improvement of such opinions by iteration of the deliberation process more than once. Corollary 4.3.1 on the other hand, allows a generalisation that makes the iteration of the deliberation process meaningful. For groups with lower second order reliabilities, the assessment of the reliabilities improve in each round of the deliberation. The iteration of the deliberation process will thus improve these second order reliabilities until the assumption of Corollary 4.3.1 is satisfied and the emergence of convergence is guaranteed.

4.3.1 Comparison with Majority Voting

We will now compare our deliberation model with majority voting. Let $\mathcal{X} = \{(\pm V_1, \ldots, \pm V_n) \mid +V_i = V_i, -V_i = \neg V_i\}$ be the set of all possible voting profiles for a group of size n. A decision rule on $\mathcal{X}$ is a function $f : \mathcal{X} \to \{V, \neg V\}$, that for each voting profile returns a (collective) vote for the hypothesis. As argued in details in (Dietrich 2006) the epistemically optimal decision rule is the weighted average where the weights are given by the likelihood ratios. For

homogeneous groups this weighted average is reduced to simple majority voting as all group members have the same likelihood ratio and thus the same weight in the averaging process. For groups with very high second order reliabilities the estimated likelihood ratios correspond to the correct values and as one can notice from Theorem 4.3.1, for reliable homogeneous groups, the deliberation process will result in a group consensus on the correct (respectively, wrong) answer if and only if the majority of group members vote correct (respectively, wrong) initially. By the same theorem the subjective beliefs will stabilise on the true belief (respectively, wrong belief) if and only if the majority of group members vote correctly (respectively, wrongly) in the beginning. Thus:

Proposition 4.3.2 *For a reliable homogeneous group G with high second order reliabilities, the deliberation process has no epistemic advantage to majority voting and vice versa.*

The advantage of the deliberation process for these groups, however, is that the group will arrive at a consensus and all group members agree on the collective decision. This is in contrast to majority voting where a minority has to accept the resulting compromise without actually endorsing it. Hence, the advantage of deliberation to majority voting for these groups is merely procedural. For unreliable homogeneous groups, however, the deliberation process comes with *some* epistemic advantage. For these groups the majority voting is doomed to end with the wrong choice for large groups by the same argument as in the Condorcet Jury Theorem. The deliberation process, however, may converge to the correct answer (depending on the group size and the initial probabilities). For groups with lower second order reliabilities, however, one would expect the majority voting to preform better than the deliberation procedure.

4.4 Inhomogeneous Groups

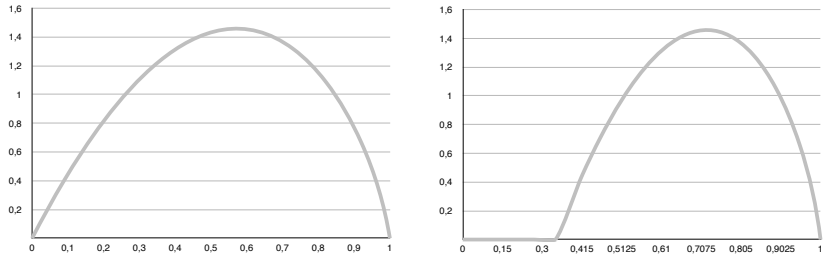
In this section we will use computer simulations to investigate inhomogeneous groups with second order reliabilities less than 1. In Section 4.4.1 we will see that the simulation results suggest that the deliberation process correctly tracks the truth in these cases as well. We shall also present an illustration of Theorem 4.3.2 and the argument preceding it. Finally, in Section 4.4.2 we will explore which of the two procedures – voting and deliberation – performs better on epistemic grounds.

Recall that in the deliberation procedure, each group member a_i has to estimate the reliability of her fellow group members a_j $(j \neq i)$ and to assign a corresponding value to r_{ij}. To determine these values we use the (initial) second order reliabilities $c_i^{(0)} \in (0,1)$. Group members with a high value of $c_i^{(0)}$ give a

more accurate assessment of r_j's. To model this, we assume that the reliability $r_{ij}^{(0)}$ is calculated from a β-distribution translated to an interval around r_j. The length of this interval is defined by the $c_i^{(0)}$. Higher values of $c_i^{(0)}$ will result in smaller intervals surrounding r_j and thus a more accurate estimation. To do so we consider a β-distribution with parameters[6]

$$\alpha = 2 \quad , \quad \beta = \frac{\min(1, r_j - c_i^{(0)} + 1) - \max(0, r_j + c_i^{(0)} - 1)}{r_j - \max(0, r_j + c_i^{(0)} - 1)}$$

in $[0, 1]$ which is then linearly transferred to the interval $\left[\max(0, r_j + c_i^{(0)} - 1), \min(1, r_j - c_i^{(0)} + 1)\right]$. See Figure 1. The values α and β are set such that the β-distribution has the mode r_j after it is transferred to the required interval.



(a) β-distribution on (0 , 1)

(b) β-distribution in (a) transferred to the open interval $(0.35, 1)$

Figure 4.1: The β-distribution with parameters $\alpha = 2$ and $\beta = 1.625$ corresponding to $r_j = 0.75$ and $c_i = 0.6$.

We furthermore assume that the group members become more competent in estimating the reliabilities of others. Thus, in each round, we also update the estimated reliability values so that they come closer to the objective values. That is we recalculate the estimated reliabilities from a β-distribution transferred to a smaller interval defined by an updated value of the second order reliabilities. More specifically, we assume that the second order reliability $c_i^{(k)}$ in round k increases linearly as a function of the number of rounds until a maximum value

[6]It turns out that our results do not vary much with the value of α. What counts is that the β-distribution has the mode r_j after it is transferred to the interval defined by r_j and $c_i^{(0)}$.

$C_i \leq 1$ is reached after M rounds. Afterwards, $c_i^{(k)}$ remains constant. Hence,

$$c_i^{(k)} = \begin{cases} (C_i - c_i^{(0)}) \cdot k/M + c_i^{(0)} & : \quad 0 \leq k \leq M \\ C_i & : \quad k > M \end{cases} . \tag{4.8}$$

4.4.1 Truth Tracking

Figure 4.2, shows the probability of tracking the truth in the deliberation as a function of group size. We examine inhomogeneous groups with unreliable members comprising the minority (Figure 4.2 (a)) and the majority (Figure 4.2 (b)) of the group members. As the simulation results suggest, in both cases the deliberation tracks the truth for large group sizes. Notice that the group in Figure 4.2 (b) has an average reliability of less than 0.5 but given the low second order reliabilities the group members do not have access to each others correct likelihood and only estimate these values in a rather large interval.

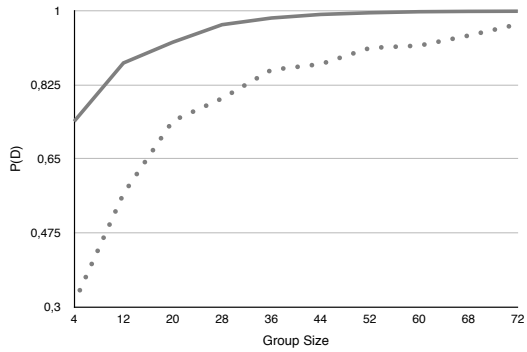


Figure 4.2: P_D for inhomogeneous groups as a function of the group size. (a) 1/4 of the group members has a reliability of 0.25, the rest has a reliability of 0.7 (solid line). (b) 1/4 of the group members has a reliability of 0.7, the rest has a reliability of 0.25 (dotted line).

In Figure 4.3 we consider an unreliable homogeneous group. As we argued above the probability of reaching a consensus on the correct answer can oscillate as the group moves from the case where the majority vote correctly to the case where the majority vote incorrectly.

We conclude that the deliberation procedure (as modelled above) is truth-conducive under similar conditions that hold for the Condorcet Jury Theorem, but we note that we do not have an analytical proof for this. The statement is only suggested by the results of our simulations.

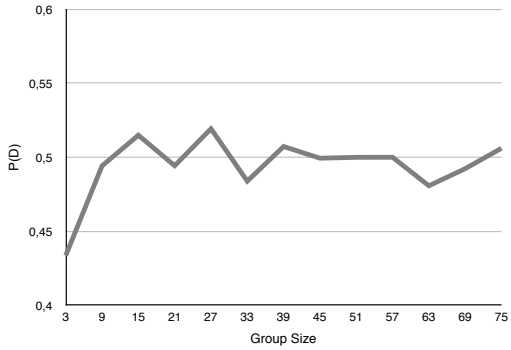


Figure 4.3: P_D for a homogeneous group as a function of the group size. Each group member has a reliability of 0.4.

4.4.2 Comparison with Majority Voting

We have already argued that the deliberation process presents no epistemic advantage over majority voting for homogeneous groups with high second order reliabilities and that for homogeneous groups with low second order reliabilities majority voting does better than our deliberation procedure. Let us now compare both procedures for various inhomogeneous groups.

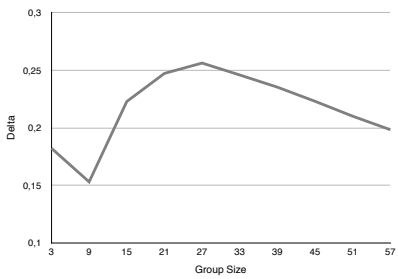
In what follows, let P_D and P_V denote the probability of converging to the correct result through deliberation and voting respectively and let

$$\Delta = P_D - P_V.$$

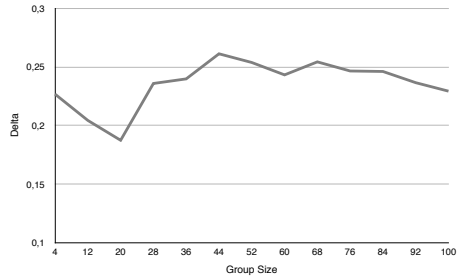
Unless otherwise stated, we plot Δ as a function of the group size n. Unless expressed differently, in all the simulations the second order reliability of the group members start from 0.6 and is increased linearly, notice that the second order reliability of 0.6 defines an interval of maximum length 0.8 centred around each r_j (cut at zero or one when necessary) thus allowing for possibly very inaccurate estimations. To control the noise, we set the number N of simulations to 10^5 and in some case to 10^6.

In Figures 4.4(a) and 4.4(b), the majority of the group members (2/3 and 4/5, respectively) has a high reliability and the rest has a low reliability. In Figures 4.4(c) and 4.4(d) the situation is reversed while in all cases the average reliability is above 0.5. The simulation results suggest that for inhomogeneous groups the deliberation procedure shows epistemic advantage over majority voting. The difference, however, is more visible for small and medium size groups and becomes smaller as the size of the group increases.[7] Figure 4.5 shows the comparison

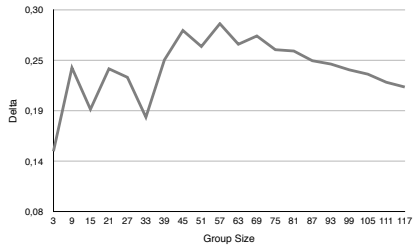
[7]This is, of course, not surprising as P_V (*pace* Condorcet Jury Theorem) and P_D (as sug-



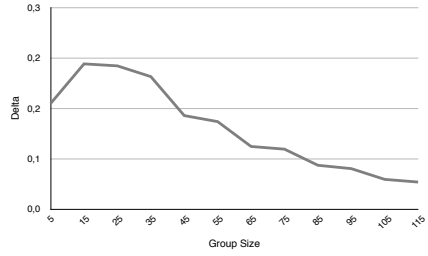
(a) 2/3 of the group has a reliability of 0.7, the rest has a reliability of 0.25.



(b) 3/4 of the group has a reliability of 0.6, the rest has a reliability of 0.35.



(c) 1/3 of the group has a reliability of 0.8, the rest has a reliability of 0.4.



(d) 1/5 of the group has a reliability of 0.95, the rest has a reliability of 0.45.

Figure 4.4: Δ as a function of the group size.

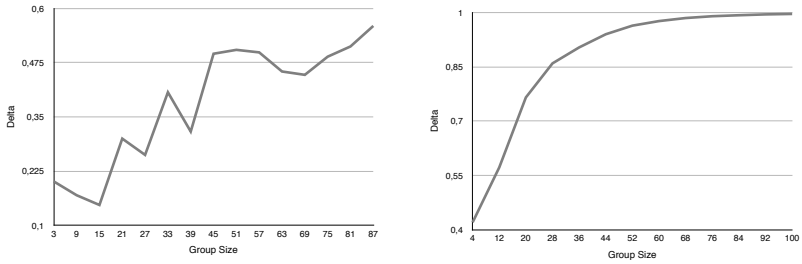
between the deliberation procedure and majority voting for two inhomogeneous groups with average reliabilities of less than 0.5.

The comparison of the deliberation procedure and the voting procedure also depends the second order reliabilities. The probability of the correct choice in deliberation is positively correlated with the second order reliabilities while voting depends only on the first order reliabilities. Thus the difference between deliberation and voting increases for the higher values of second order reliabilities and decreases for lower values.

Figure 4.6 shows the difference between truth tracking in deliberation and voting as a function of the (initial) second order reliability for three different groups sizes ($n = 15, 27$ and 33) with the same distribution of (first order) reliabilities: 2/3 of the group has reliabilities of 0.6 and the rest has reliabilities of 0.75.

As the graph suggests, the result of the comparison depends highly on the

gested by our simulations) coverage to 1.



(a) 1/3 of the group has a reliability of 0.75, the rest has a reliability of 0.35.

(b) 1/4 of the group has a reliability of 0.8, the rest has a reliability of 0.25.

Figure 4.5: Δ as a function of the group size.

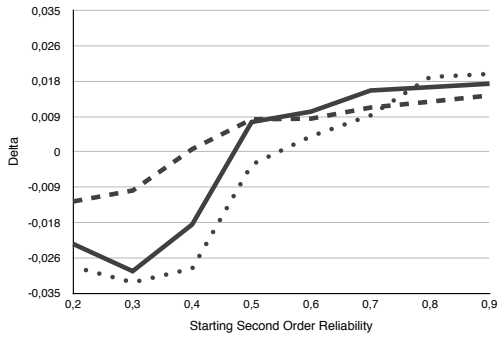


Figure 4.6: Δ as a function of the (initial) second order reliability for different group sizes n: $n = 15$ (dotted line), $n = 27$ (solid line), and $n = 33$ (dashed line).

(initial) second order reliabilities. Initial second order reliabilities greater than 0.6, 0.5 and 0.4 make the deliberation procedure epistemically better for groups of size $n = 15, 27$ and 33, respectively, while for lower (initial) second order reliabilities the voting procedure performs better.

Finally, Figure 4.7 shows a group with one highly reliable member where the other group members have near average reliabilities.

4.5 Conclusions

Voting and deliberation are two standard procedures to reach a group decision. The goal of this paper was (i) to present a new Bayesian model for non-strategic rational deliberation, (ii) to study the emergence of consensus and its truth track-

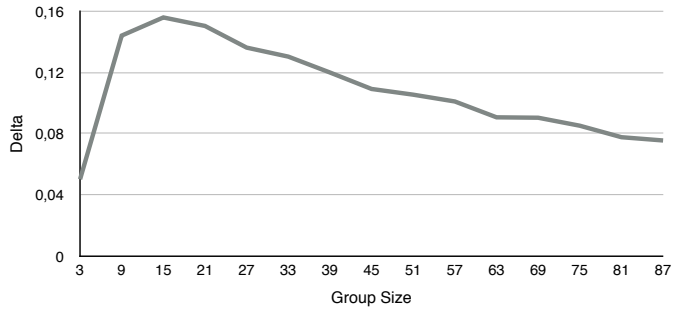


Figure 4.7: Δ for a group with only one highly reliable member. One member has a reliability of 0.9, the rest has a reliability of 0.55. The (initial) second order reliability is 0.85.

ing properties, and (iii) to compare this deliberation process with majority voting in terms of their truth-tracking properties. To this end, we proposed a Bayesian model which allows for such a comparison. The model is based on two attributes of the group members: we assumed that each group member has a (first order) reliability to make the right decision, which equals the Condorcet reliability, and a second order reliability to assess the (first order) reliability of the other group members. We furthermore assume that each group member updates her probability that the hypothesis is true in each deliberation round based on the previous verdicts of the other group members.

We have shown that the deliberation process results in a consensus and correctly tracks the truth for groups of large size inn the following cases: (i) homogeneous groups with a first order reliability greater than 0.5 and with a high second order reliability. (ii) inhomogeneous groups with average reliabilities above 0.5 and with a high (initial) second order reliability. In this sense the deliberation procedure manifests the same epistemic properties as the majority voting while adding the benefit of a group consensus which for groups with average reliabilities above 0.5 and high (initial) second order reliabilities will make sure that all group members reach a stable correct belief about the hypothesis in finitely many steps. We furthermore provided some simulation results that indicate that the deliberation procedure tracks the truth even in cases that do not fall under the conditions stated in the Condorcet Jury Theorem for majority voting as well as for groups with low second order reliabilities.

Clearly, these results are consequences of our assumptions. But how robust are the results? Do they also hold if we make changes in our deliberation model

and relax some of its idealisations? In future work, we would especially like to study the effect of relaxing the independence assumption. While it makes sense for voting, the independence assumption is questionable for deliberations as more and more links between the group members are established in the course of deliberation which make the group members (and henceforth also their verdicts) dependent on each other. At the end of the deliberation process, when a consensus is reached, it is as if the original assembly of independent individuals has become one homogeneous entity, with all group members endorsing the consensus. The challenge, then, is to model how a social network emerges in the course of the deliberation process and to explore how it becomes (under conditions to be explored) increasingly dense as the process proceeds.

Other issues that require further work concern the investigation of groups with low second order reliabilities and the study of mechanisms for updating the second order reliabilities in the course of deliberation. The justification for updating the second order reliabilities is that through the course of deliberation group members will get a chance to evaluate each other's arguments and form a better judgment of each other's reliability. But this would also suggest that those with higher first order reliabilities are in a better position to judge the validity of others' arguments. As such the second order reliabilities of these members should increase faster than those with a lower first order reliability. This suggests that the process of updating the second order reliability of a group member should take her first order reliability into account.

4.6 Appendix

A1. Proof of Theorem 4.3.1

First notice that since all group members have the same reliability $r_i = r$ and the same second order reliability $c_i = 1$, the estimated reliabilities in each round will be equal to the actual reliabilities and the likelihood ratio will be the same for all group members in each round, i.e. $x_{ij}^{(k)} =: x = (1 - r_{ij}^{(k)})/r_{ij}^{(k)} = (1 - r)/r$. So

$$\begin{aligned} P_i^{(k+1)}(\mathrm{H}) &= P_i^{(k)}(\mathrm{H}|\mathrm{Vote}_1^{(\mathrm{k})}, \ldots, \mathrm{Vote}_{\mathrm{i-1}}^{(\mathrm{k})}, \mathrm{Vote}_{\mathrm{i+1}}^{(\mathrm{k})}, \ldots, \mathrm{Vote}_{\mathrm{n}}^{(\mathrm{k})}) \\ &= \frac{P_i^{(k)}(\mathrm{H})}{P_i^{(k)}(\mathrm{H}) + (1 - P_i^{(k)}(\mathrm{H}))\ \prod_{j\neq i=1}^{n} \left(x_{ij}^{(k)}\right)^{p_j^{(k)}}} \\ &= \frac{P_i^{(k)}(\mathrm{H})}{P_i^{(k)}(\mathrm{H}) + (1 - P_i^{(k)}(\mathrm{H}))\, x^{\sum_{j\neq i=1}^{n} p_j^{(k)}}}, \end{aligned} \tag{4.9}$$

where $p_j^{(k)} \in \{0,1\}$ is the vote of group member a_j in round k and $p_j^{(k)} = 1$ if $Vote_j^{(k)} = V_j$, i.e. if group member a_j has voted (correctly) for the truth of the hypothesis and $p_j^{(k)} = -1$ otherwise. Simplifying this we have $P_i^{(k+1)}(\mathrm{H}) > P_i^{(k)}(\mathrm{H})$ if and only if $x^{\sum_{j\neq i=1}^{n} p_j^{(k)}} < 1$.

The votes in the first round are given by the initial probability assignments that arise from the group members' reliabilities r. This means that group member a_j will start by initially voting correctly, i.e. $p_j^{(0)} = 1$ (or equivalently $P_j^{(0)}(\mathrm{H}) \geq 0.5$) with probability r and incorrectly, i.e. $p_j^{(0)} = -1$ (or equivalently $P_j^{(0)}(\mathrm{H}) < 0.5$) with probability $1-r$. Thus $P_i^{(1)}(\mathrm{H}) > P_i^{(0)}(\mathrm{H})$ if and only if $x^{\sum_{j\neq i=1}^{n} p_j^{(0)}} < 1$. Since $r > 0.5$ and $x < 1$, $x^{\sum_{j\neq i=1}^{n} p_j^{(0)}} < 1$ if and only if $\sum_{j\neq i=1}^{n} p_j^{(0)} > 0$ that is if the majority of the group members (excluding a_i) vote correctly in the first round.

Notice that if the majority of the group members votes correctly in some round, say in round t, and if $p_i^{(t)} = -1$ then the majority of the group excluding a_i has voted correctly in round t and thus $\sum_{j\neq i=1}^{n} p_j^{(t)} > 0$. If, however, $p_i^{(t)} = 1$ it is possible that $\sum_{j\neq i=1}^{n} p_j^{(t)} = 0$ that is when there are exactly the same number of correct and incorrect votes in the rest of the group. In this later case $P_i^{(t+1)}(\mathrm{H}) = P_i^{(t)}(\mathrm{H})$. However, since the probability assignment for any member who has voted incorrectly in round t strictly increases, after some finite number of rounds, say l, the probability assignment for at least one of these group members, say a_s, will increase enough such that $p_s^{(t+l)} = 1$ and from then on we have that the number of correct votes in the whole group is at least two more than the number of incorrect ones and thus $\sum_{j\neq i=1}^{n} p_j^{(t+l)} > 0$ for $i = 1, \ldots, n$. Thus for simplicity of notation and without loss of generality we can assume that when the majority of the group votes correctly initially, the number of correct votes is at least two more than the number of incorrect votes. Thus $\sum_{j\neq i=1}^{n} p_j^{(0)} > 0$ for $i = 1, \ldots, n$ and so $P_i^{(1)}(\mathrm{H}) > P_i^{(0)}(\mathrm{H})$ for $i = 1, \ldots, n$. Similarly when we consider the case where the majority of the group members vote incorrectly in the first round we shall assume that the number of incorrect votes is at least two more than the number of correct ones.

In the second round of the deliberation the votes will be casted on the basis of the updated probability assignments. Thus if $P_i^{(1)}(\mathrm{H}) > P_i^{(0)}(\mathrm{H})$ for $i = 1, \ldots, n$ then $\sum_{j\neq i=1}^{n} p_j^{(1)} \geq \sum_{j\neq i=1}^{n} p_j^{(0)} > 0$ since each group member j who had voted for the truth of the hypothesis on the basis $P_j^{(0)}(\mathrm{H})$ will still vote the same on the basis of the equal or higher probability $P_j^{(1)}(\mathrm{H})$ while some of the group members who had voted against the hypothesis may change their vote if their subjective probability has been raised to a value above 0.5. Hence from $\sum_{j\neq i=1}^{n} p_j^{(1)} > 0$ we

have $P_i^{(2)}(\mathrm{H}) > P_i^{(1)}(\mathrm{H})$ for $i = 1, \ldots, n$.

Repeating the same argument the subjective probabilities of the group members (for the truth of the hypothesis) will increases in each round and will be greater or equal to 0.5 in finitely many steps. Thus if the majority of the group members vote correctly in the first round the group will reach a consensus on the correct answer in finitely many steps. If the group members keep repeating the deliberation process (possibly even after the consensus is reached) the probabilities will increase until at some round t, we have $P_i^{(t)}(\mathrm{H}) = 1$ for $i = 1, \ldots, n$ after which repeating the deliberation process will no more change the probabilities. This proves part (ii).

By the same argument, if the majority of the group members vote incorrectly in the first round the probability assignments will decreases until after finitely many steps all group members will assign probability zero to H and the group will reach a consensus and the subjective beliefs will stabilise (on the wrong belief) and this gives the result for part (iii). Parts (ii) and (iii) will together imply part (i), as it is either the case that the majority have voted correctly in the first round or that the majority have voted incorrectly and in either case the group will reach a consensus in finitely many rounds (on the correct answer and incorrect answer respectively).

If $r \geq 0.5$ then by the Condorcet Jury Theorem the probability that the majority of the group members would vote correctly in the first round (and thus the group reaches a consensus on the correct answer), increases with the size of the group and approaches 1 as the size of the group increases. Similarly if $r < 0.5$ by the same argument as in the Condorcet Jury Theorem the probability that the majority of the group members would vote incorrectly in the first round (and thus the group reaches a consensus on the wrong answer), increases with the size of the group and approaches 1 as the size of the group increases. This proves part (i).

A2. Proof of Theorem 4.3.2

Since $r < 0.5$ and thus $x > 1$, by the argument in the proof of Theorem 4.3.1, if the majority of the group members start by voting incorrectly we have that $\sum_{j \neq i=1}^{n} p_j^{(0)} < 0$ and thus $x^{\sum_{j \neq i=1}^{n} p_j^{(0)}} < 1$ and the probability assignments increases until the majority will assign a subjective probability above 0.5 to hypothesis at some round t (and thus vote correctly) after which $x^{\sum_{j \neq i=1}^{n} p_j^{(t)}} > 1$ and the subjective probabilities will decrease and this will repeat. Similarly if the majority start by voting correctly the subjective probabilities will decrease until at some stage the majority will assign a probability less than 0.5 to the hypothesis after which they will vote incorrectly and thus the probability assignments will start to increase, etc.

Chapter 5

Anchoring In Deliberations

5.1 Introduction

There are numerous instances of group decision making in everyday practice. Families have to decide where to go on holiday, funding agencies have to decide which research projects to support, and juries in court have to decide whether a defendant is guilty or not. Sometimes a decision is made in the light of different preferences (and each group member wants to get the best out for herself), and sometimes all group members share the conviction that the resulting decision should be best in some sense that is commonly agreed upon. Juries in court, for example, want to make the right decision. Everybody wants that a guilty person is convicted, and no one wants that an innocent is sent to prison. Likewise, committees such as the IPPC deliberate to arrive at the best far-reaching policy recommendations.

There are many different ways that a group can make such decisions. When deciding in light of different preferences, for example, the goal is naturally to minimise the average compromise and leave the decision makers as happy as possible with the final decision and when deciding toward some objective truth the goal is to approximate the correct decision as close as possible. Different decision making process are, thus, chosen based on which characteristics they manifest the best.

Deliberation is one extensively practiced approach to decision making in both such scenarios. It provides an advantage of allowing the decision makers to revise and change their opinion and thus stabilising a dynamics that can ideally lead the group to a collective consensus on which the group agrees as a unanimous entity. This is indeed a sought after characteristic for a group decision making procedure to leave all the participant in agreement and convinced of the final

decision. On the other hand, the possibility of learning other opinions paves the way for easier recognition of mistakes and better availability of information which help to better estimate the correct decision when such exists.

Adding the interactive component to the decision making, however, does also introduce a whole spectrum of other relevant factors, including those associated with the psychological aspects of social interactions. There are many factors that can affect how people influence each other in their interactions, from the way that one reacts to the opinions of those from different social, economical or educational classes to the intricate and complex dynamics of power in social networks. Phenomena resulting from such factors, including *information cascades*, *pluralistic ignorance*, *anchoring*, etc., have been studied by psychologist for a rather long time to varying degrees. Some are well understood, analysed and agreed upon but most have proved hard to characterise in an uncontroversial way or even harder to do so in a manner exact enough to be account for in technical formalisations. Nevertheless, there is an increasing tendency in social scientists and psychologists toward development and adoption of formal models, and to our opinion, very rightly so.

This does not ,by any means, intended to suggest that we believe that the subtlety of the psychological (and socio-psychological) phenomenons can be completely and adequately captured in some mathematical and formal apparatus. Rather it is meant to emphasise on the fact that such formalisations, even with the necessary idealisations and abstractions that are inevitable in devising formal models, can surface persisting patterns and trends and shed light on general characteristics, aspects, and correlations between the relevant factors to be studied in more details (in possibly non-formal approaches). In this regard, the newly established links between the Bayesian Epistemology, as our best theory of graded belief, and the concepts in psychology (and socio-psychology) should be celebrated and carefully attended.

One such psychological phenomenon that is relevant to interaction of group members and the formation of a collective consensus is the *anchoring* effect. This is a social-psychological instance of an effect that is widely discussed in the heuristics and biases program in cognitive psychology, (Tversky & Kahneman 1974), (Kahneman et. al. 2006). In general anchoring is a cognitive bias that describes the common human tendency to rely heavily on some (usually irrelevant) piece of information when making decisions.

There are many different instances of anchoring effect arising even in an individual decision making. As it relates to the deliberation process, anchoring occurs when the group consensus depends, for example, on the order in which the group members present their views in the course of a deliberation. More specifically, the experiments suggest that the group member who speaks first will typically have the highest impact on the decision on which the group eventually settles. In

such cases the first speaker is said to have *anchored* the deliberation. The effect is particularly important for the epistemic analysis of the deliberation as it opens the way to possible manipulation of the process and thus bears negatively on the epistemic characteristic of the final decision.

The reasons for the emergence of the anchoring effect are usually associated with what is known as the *bounded rationality*. These are the cognitive limitations of the decision makers including, short attention span, memory loss, loose of cognitive ability by fatigue, etc. That the effect can happen as a result of such limitations seems clear. So the question we will ask here is whether or not such cognitive limitations are the only causes for this bias? In other words we ask whether the anchoring effect can also occur in a group of truth-seeking, fully rational members, who update their beliefs according to plausible rules, simply as a result of the updating procedure.

To address this question, the remainder of this paper is organised as follows. We shall first proposes an incremental model for deliberation, inspired by the well-known Lehrer-Wagner model in Section 5.2. Using this model, we study the anchoring effect in Section 5.3 and will conclude in Section 5.4.

5.2 Modeling Anchoring

To study the emergence of the anchoring effect, we will first need a formalisation of the deliberation process. There are several models of deliberation studied in the literature including two Bayesian models developed by Angere and Olsson, (3), (Olsson 2011), and Hartmann and Rafiee Rad (Chapter 4). Both these models, however, make a crucial independence assumption by which the order in which the evidence comes in will be inconsequential. Thus it will not be possible to study the path-dependence of the deliberation process with these models.

Next are the Lehrer-Wagner model, (Lehrer 1976), (Lehrer & Wagner 1981), and Hegselmann-Krause model (Hegselmann & Krause 2002; Hegselmann & Krause 2006; Hegselmann & Krause 2009). The Lehrer-Wagner model is probably the most well known model in the literature for collective consensus which focuses on deliberating groups that have to fix the value of a real-valued parameter. According to the model, each group member submits her initial assignment and assigns normalised weights to all group members, including herself. The revised assignment of each group member will be the weighted average of all initial submissions using the assigned weights. It is not hard top show that, if the process is iterated, under rather weak conditions the values will converge. This model, however, assumes that all assignments in each round are made simultaneously, and so again the supposed path-dependence of the deliberation cannot be accounted for in this model. The Hegselmann-Krause model has a similar problem. According to this model, each agent takes into

account only those judgments that are sufficiently close to her own initial assignment and the updating procedure happens on the full profile of all such opinions at the same time.

It is important to note, however, that these models are all intended as normative models for rational deliberation developed in such a manner to avoid such biases as the anchoring. To study the emergence of this effect we will need a descriptive model for deliberation or one *closer* to how the actual deliberations are carried out in groups. In real deliberations the group members are hardly independent, and even if they start as such, during the course of the deliberation dependencies will form and become more and more complicated as the process goes on. It is also usually not the case that people would wait to hear all the opinions before updating theirs and every opinion and argument does leave some influence as it is presented. We will next propose a simple formalisation of the deliberation process (a modified version of the Lehrer-Wagner model) that captures this intuition and allows fro the study of path-dependence in deliberation. As in the Lehrer-Wagner model we consider a group aiming to estimate the value of a real valued parameter x through deliberative process. The model works on the following premises:

- Each group member is assumed to have a first order reliability that captures his competence in giving the correct judgement and a second order reliability that captures his ability to estimate the competence of his fellow group members.

- Each group member estimates the reliability of her fellow group members and use these estimated reliabilities to weight the opinions of others.

- The updating procedure proceeds in an incremental manner, that is, the group members update their opinion after each announcement. Thus each round of deliberation in a group of n agents, consists of n steps. In each step a group member announces her opinion and everyone updates based on this announcement.

- The way each group member updates her opinion depends on the reliability that she assigns to speaker in relation to the reliability she assigns to herself.

- People's ability to estimate the competence of others improve during the course of deliberation.

So here is the sketch of how the model works; Consider a group of n members who have to fix the value of a real-valued parameter x and let's order the group members from 1 to n. Initially, each group member i submits an initial value $x_i^{(0)}$ and fixes a value for the reliability of her judgment. In round 1, step 1, the first

group member presents her arguments for her assignment (i.e. for $x_1^{(0)}$). Based on this, the other group members (i) assign a reliability to her and then (ii) update their original submissions. Thus the i^{th} group member updates her initial value (i.e. $x_i^{(0)}$) taking $x_1^{(0)}$ as well as her own reliability and the reliability she assigns to the first group member into account. In the next step, the second group member presents her arguments for her (now already once updated) assignment, and group members $1, 3, \ldots, n$ update their assignments based on this and so on. In the second round of deliberation, the same procedure repeats. We assume, however, that the reliability assignments improve, i.e. that the group members become increasingly better in judging the reliability of their fellow group members.

Let's now formalise this procedure. To do so, we fill in the details for the updating procedure and the process of estimating the reliabilities.

5.2.1 The Estimation of Reliabilities

We assume that every group member i (for $i = 1, \ldots, n$) has an objective reliability r_i. While these are real numbers in $(0, 1)$, the reliabilities that are actually used by the group members have discrete values. More specifically, we assume (for simplicity) that there are only three possible reliability values: H (high), M (medium) and L (low). We furthermore assume that every group member i has access to her own objective reliability and assigns herself a reliability of H if $r_i \geq 2/3$, L if $r_i \leq 1/3$ and M otherwise.[1]

Next, The group members estimate each others' reliabilities. To do so, we assume that everybody has a second order reliability $c_i \in (0, 1)$ that captures their competence in assessing the reliability of others. Individuals with a high value of c_i give a more accurate assessment of r_j's (reliabilities of other group members) than those with a low value of c_i. To be more precise we assume that the group member i estimates the reliability of the group member j, r_{ij}, from a β-distribution on an interval with the length $2(1 - c_i)$ around r_j.

$$r_{ij}^{(0)} = \beta - \text{Distribution}\left[\max(0, r_j + c_i - 1), \min(1, r_j - c_i + 1)\right] \tag{5.1}$$

Thus, for $c_i = 0$, the β-distribution extends over the whole interval $(0, 1)$. And for $c_i = 1$, we obtain $r_{ij}^{(0)} = r_j$. Using these estimated reliabilities, each group member i, assigns an effective reliability to the group members j: H if $r_{ij} \geq 2/3$, L if $r_{ij} \leq 1/3$ and M otherwise.We also assume that the estimation of the reliability improve in each round of the deliberation. That is, in each round,

[1] This is a strong assumption and can of course be relaxed. Indeed, psychological evidence suggests that people are not good at assessing their own reliabilities. Notice however that since the group members use only the reliability brackets High, Medium and Low, we only need to assume that each group member has access to her reliability bracket as opposed to exact value of her objective reliability.

we also update the estimated reliability value so that they come closer to the actual values. More specifically, we assume that the competence $c_i^{(k)}$ in round k increases linearly as a function of the number of rounds until a maximum value $C_i \leq 1$ is reached after K rounds. Afterwards, $c_i^{(k)}$ remains constant. Hence,

$$c_i^{(k)} = \begin{cases} (C_i - c_i) \cdot k/K + c_i & : \quad 0 \leq k \leq K \\ C_i & : \quad k > K \end{cases} . \tag{5.2}$$

Note that updating the reliabilities justifies that the deliberation process proceeds in several rounds. In each round, the group members learn something more about the reliability of their fellow group members. And this is why they will keep on updating. However, it seem natural to stop the updating procedure after some finite number of rounds, say K. Clearly, the value of K will depend on contextual factors such as how patient the individuals are. If no consensus is reached after round K, then the straight average will be taken.

5.2.2 The Updating Procedure

We should now define the updating procedure. In step 1 of round 1, group member 1 presents her arguments. In this step, she does not change her original submission, and she does not change her own reliability assignment. All other group members update their original assignments and their own reliability assignment according to the following rules that are inspired by Elga's, (Elga 2007), discussion of reflection and disagreement. Presented in first person perspective, the updating proceeds as follows,

1. I am H (M, or L)[2] and the presenter is my *peer*, i.e. she is also H (M, or L). In this case, my new assignment is the straight average of her assignment and mine:
$$x_i^{(1)} = \frac{1}{2}\left(x_i^{(0)} + x_1^{(0)}\right).$$
My reliability value remains H (M, or L).

2. I am H, the presenter is L. In this case I disregard the opinion of the presenter and stick to my original judgment:
$$x_i^{(1)} = x_i^{(0)}.$$
My reliability remains H.

[2]For ease of writing, we assume that I (i.e. group member i) am one of the other group members, i.e. "I am H" is short hand for "My reliability value is H" etc.

3. I am L, the presenter is H. In this case I accept the opinion of the presenter:
$$x_i^{(1)} = x_1^{(0)}.$$
My reliability changes to H.

4. I am H (M), the presenter is M (L). In this case, my new assignment is the weighted average of her and my original assignment:
$$x_i^{(1)} = \frac{1}{4}\left(3x_i^{(0)} + x_1^{(0)}\right).$$
My reliability value remains H (M).

5. I am L (M), the presenter is M (H). In this case, my new assignment is the weighted average of her and my original assignment:
$$x_i^{(1)} = \frac{1}{4}\left(x_i^{(0)} + 3x_1^{(0)}\right).$$
My reliability value changes to M (H).

Rule 1 expresses the Equal Weight View, (Elga 2007). Rules 2 and 3 are inspired by Elga's discussion of the guru case. Rules 4 and 5 use weights that reflect that the reliability assignments in question are one step apart from each other. Note however that the exact values of the weights in Rules 4 and 5 are not important and they have been chosen for the ease of calculation. The only crucial point here is to assign a higher weight to the opinion that corresponds to the higher reliability bracket. In step 2, group member 2 presents and the other group members update according to the above rules and so on. After n steps, every group member has presented once and round 1 is over. Perhaps a consensus is already reached. If not, the group might decide to deliberate for a second or third round.

Although the model introduced here is overly simple it does indeed provide the necessary setting for our study in so far that it allows formation of inhomogeneous groups with different weights of opinion and an incremental updating procedure that reflects these differences.

5.3 The Anchoring Effect in Deliberations

We shall study two types of grouse separately. The *homogeneous group* where the group members consider each other as epistemic peers and *inhomogeneous groups* where the group members have different reliabilities. For the first case we shall give analytical results for the emergence of anchoring and for the second case we will use computer simulations and numerical analysis.

5.3.1 Homogeneous Groups

There are generally two approaches to how one should take the opinion of her epistemic peers into consideration. The first approach advocates assignment of equal weights (or almost equal weights) as in the updating procedure we shall use here. The second, requires one to hang on to her own belief and refrain from any updating in case of a disagreement with epistemic peers. The latter, however, does not go well with the spirit of deliberation. Refusing to update one's opinion in case of disagreement among epistemic peers will prevent such groups from meaningful deliberation and reaching a consensus. With the former approach, in a group of epistemic peers, after the i^{th} speaker announces his value the group members will all consider this opinion by giving it the same weight as they give to their own and thus simply averaging their assignment with the one announced by the speaker. To formalise this, if the current values estimated by the group members for the value of the parameter in question is given by $\vec{V}$, the updated values after the i^{th} speaker's announcement will be

$$\vec{V}' = 1/2 B^i_n . \vec{V},$$

$$B^i_n = A^i_n + I_n,$$

where A^i_n is an $n \times n$ matrix with 1's on the i^{th} column and zeros else where and I_n is the unit $n \times n$ matrix. In this fashion the values in V' will be the average of values in V and the announcement of the i^{th} speaker, that is $< \vec{V} >_i$. So, starting from the initial assignments given by $\vec{V}^{(0)}$, the result of one round of deliberation will be given by

$$\vec{V}^{(1)} = \frac{1}{2^n} \left(B^n_n . B^{n-1}_n . \ldots . B^1_n \right) . \vec{V}^{(0)},$$

and after k rounds,

$$\vec{V}^{(k)} = \frac{1}{2^{nk}} \left(B^n_n . B^{n-1}_n . \ldots . B^1_n \right)^k . \vec{V}^{(0)}.$$

To show the anchoring effect we should show that the initial opinion of the first speaker receives a higher weight (in $\vec{V}^{(k)}$) compared to other group members. We should note that this would still be the case even if the the group members use a non-equal but uniform assignment of weights. So, for example, if every one assigns higher weight, say 2/3, to her own opinion and 1/3 to the speaker. With the way that the updating procedure takes place, the first speaker will inevitably receives a higher weight. We shall make this precise in the following two theorems before moving to the case of inhomogeneous groups.

Theorem 5.3.1 *The process of deliberation described above will converge to a consensus.*

Proof Notice that the matrix representation of the updating in each round of deliberation,

$$\frac{1}{2^n}\prod_{i=1}^{n} B_n^i$$

is essentially a weight matrix in the sense of Lehrer-Wagner model. The convergence (in the limit) thus follows from the same argument as for the Lehrer-Wagner model. □

Theorem 5.3.2 *In the consensus value reached by a homogeneous group through the process of deliberation described above, the opinion of the first speaker receives a higher weight than any other group member. Moreover the opinion of the i^{th} speaker receives a higher weight than all those who speak after her.*

Proof. See appendix.

We shall discuss the stability of this result in the appendix where we will show that with small deviation from the equal weight assignment, the results from Theorem 5.3.2 still holds. In that case the assignment of weights will be taken to be $\frac{1+\epsilon}{2}$ and $\frac{1-\epsilon}{2}$. Depending on whether the ϵ is positive or negative the group members can assign a higher reliability to themselves or to the speaker. (See the Appendix B for details.)

5.3.2 Inhomogeneous Groups

There are two main differences between the homogeneous and inhomogeneous groups. First, since the group members are not epistemic peers anymore they cannot consider all the group members (including themselves) as equally reliable. Instead, they need to decide their stand in relation to other group members by assigning reliabilities to themselves and to others. Second, the updating procedure will be based on these reliability assignments which will no more amount to simply averaging the values. The actual distribution of reliabilities will play a crucial role here as will the placement of the group members with higher reliabilities.

We will investigate the emergence of anchoring effect in these groups using computer simulations. We will look at the groups consisting a mixture of reliabilities and, in particular, groups in which the first speaker has lower reliability than the rest of the group as well as the groups in which there are members more reliable than the first speaker and members that are less reliable. There are of course group combinations where the emergence or non-emergence of the effect is trivial. If the first speaker is highly reliable while the rest of the group have very low reliabilities then they will all simply adopt his assignment and the

anchoring effect is inevitable. The same way the anchoring will not emerge if the first speaker is of extremely low reliability and the rest of the group are all highly reliable. Thus the interesting cases are those groups with mixed reliabilities of high, medium and/or low in such a way that a considerable part of the group will not abandon their own assignment in favour of that of the first speaker nor will they discard her opinion all together. It is in particular interesting to see the emergence of anchoring effect in groups where a large part of the group have reliabilities higher than the first speaker (but not higher enough to completely discard her opinion). For the purpose of simulations

- for the assignment of reliability r_{ij}, assigned by group member i to the group member j, we will take a β-distribution in $[0,1]$ with parameters

$$\alpha = 2 \quad , \quad \beta = \frac{\min(1, r_j - c_i + 1) - \max(0, r_j + c_i - 1)}{r_j - \max(0, r_j + c_i - 1)}$$

  these will then be translated into reliability brackets.

- we will set the second order reliabilities to 0.8 which increase in each round of the deliberation linearly until a maximum value of 0.9 is reached where they remain fixed.

5.3.2.1 The Algorithm

The general sketch of the algorithm used for simulation is as follows:

1. Fix all relevant parameters: n = number of people; r_i = reliabilities; c_i = initial second order reliabilities; N = number of runs; C= the maximum second order reliabilities; K= number of deliberation rounds.

2. Simulate the process of deliberation: (a) Initialise the prior values assigned by the individuals using a chance mechanism. (b) Calculate estimated reliabilities using the beta distribution and second order reliabilities. (c) calculate estimated reliability brackets. (d) Start deliberation steps $1, \ldots, n$: where in step i each individual updates her assignment as well as her reliability ranking based on the assignment of the i^{th} individual. (e) If a consensus is reached, stop. If there is no consensus repeat the deliberation steps $1, \ldots, n$. (f) Repeat (e) at most K times. (g) If a consensus is reached calculate the difference between the consensus value and the initial assignments. (h) Add 1 to a counter if the consensus is closest to the value assigned by the first individual. do nothing otherwise. (i) After N runs, compute the probability that the consensus is closest to the value assigned by the first individual.

To decrease the errors in the simulations the probabilities under discussion are calculated from samples of 10^5 simulations or more.

5.3.2.2 Results

In this section we will present simulations of the deliberation process for different groups. We will calculate the probability that the final consensus of the group is closest to the initial estimation of the first speaker and plot the results as a function of the group size unless stated otherwise. The second order reliabilities are assumed to be the same for all group members. Notice that a second order eligibility of 0.8 means that the estimation are made in intervals of length at least 0.2. This means that the group members can possibly assign wrong reliability brackets to their fellow group members.

The first plot shows the anchoring effect for a group in which the first speaker

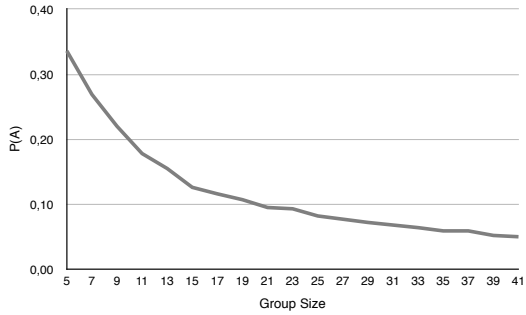


Figure 5.1: The anchoring effect as a function of the group size. First speaker's reliability of 0.85, and the rest of the group with reliabilities 0.75 and 0.4.

has reliability 0.85 and the rest of the group have reliabilities 0.75 and 0.4, equally distributed. So at least half of the group have reliabilities close to that of the first speaker. Given the assignment of reliability brackets, The first speaker will on average be assigned high reliability, half of the group will be given reliabilities high or medium and and other half will be given reliabilities medium or low. The plots shows the anchoring effect as a function of the group size.

Figures 5.2(i) and 5.2(ii) show the same group as in Figure 5.1, where the group member with high reliability (0.85) is the middle speaker and the last speaker respectively. The plots in Figure 5.2(i) and 5.2(ii) show the probability that the final consensus (if reached) is closest to the original submission of this group member (the middle speaker and the last speaker respectively).

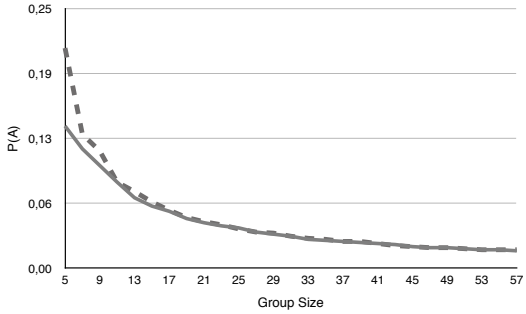


Figure 5.2: The anchoring effect for the middle and last speakers as a function of the group size. i) Dashed line for the middle speaker with reliability 0.85 and the rest 0.75 and 0.4. ii) Continuous line for last speaker with reliability 0. 0.85 and the rest 0.75 and 0.4.

Comparing Figure 5.1 with Figures 5.2(i) and 5.2(ii), shows how the anchoring effect depends on the speakers position in the group. In particular, the comparison of the plots suggests that anchoring effect depends more on the speaker's position in the group compared to her reliability as the same member placed in the middle or as the last speaker will impose a much smaller anchoring effect. To see this more clearly one can compare the anchoring effect for the first speaker in Figure 5.1 with that of the middle and the last speaker in Figures 5.3(i) and 5.3(ii) respectively, where the middle and the last speakers have significantly higher reliabilities than the rest of the group.

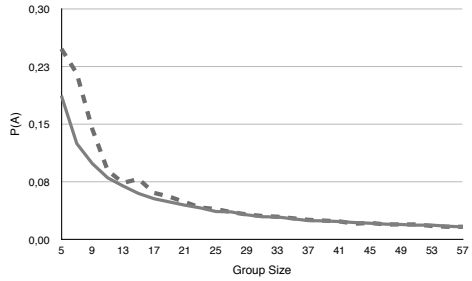


Figure 5.3: The anchoring effect for the middle and the last speakers as a function of the group size. i) Dashed line for middle speaker with reliability 0.85 and the rest 0.55 and 0.4. ii) Continuous line for last speaker with reliability 0.85 and the rest 0.55 and 0.4.

Next we will see how the first speaker's reliability can influence the anchoring effect. The first speaker in Figure 5.1 has higher reliability than the rest of the group. This, however, is not necessary for the anchoring effect and the effect emerges even when the first speaker is not particularly reliable in comparison with others. Figure 5.4 shows the anchoring effect for two groups in which all (Figure 5.4(i)) or a considerable number of group members (Figure 5.4(ii)) have reliabilities higher than that of the first speaker.

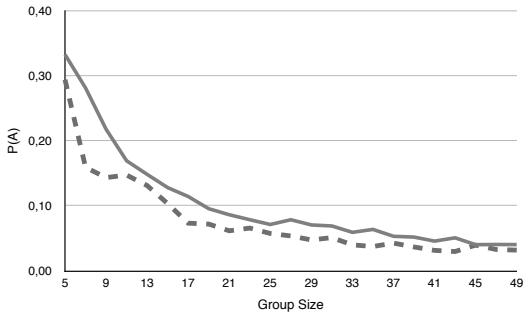


Figure 5.4: The anchoring effect for groups with first speaker less reliable than all or considerable part of the group. i) Dashed line for first speaker with reliability 0.65 and the rest 0.9. ii) continuous line for first speaker with reliability 0.75 and the rest 0.85 and 0.5.

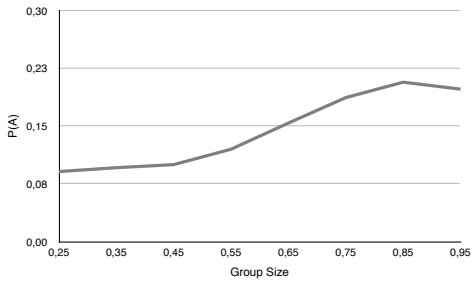


Figure 5.5: The anchoring effect as a function of the first speaker's reliability. Group size 11 with five members with reliability 0.7 and five members with reliability 0.5.

Although the anchoring effect is not particular to the first speakers with high reliability, the comparison of Figures 5.1 and Figure 5.4 suggests a positive correlation between the reliability of the first speaker and the intensity of the anchoring

effect as one would expect. To see this more clearly, we will plot in Figure 5.5 the anchoring effect as a function of the first speaker's reliability for a group of size 9 where the rest of the group members have reliabilities 0.7 and 0.5.
Finally in the next two graphs we plot the anchoring effect for two groups with random distribution of reliabilities. In the first, Figure 5.6, the reliabilities of the group members are coming from a uniform distribution in (0 1) and in Figure 5.7 the reliabilities are assigned through a β-distribution with parameters $\alpha = \beta = 2$ in (0 1). In both cases the first speaker has the reliability 0.5. In Figures 5.6 and

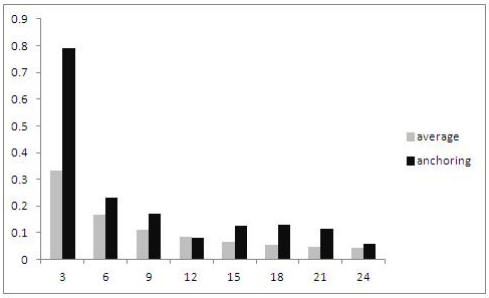


Figure 5.6: Anchoring effect as a function of the group size. Reliabilities uniformly distributed. First speaker 0.5.

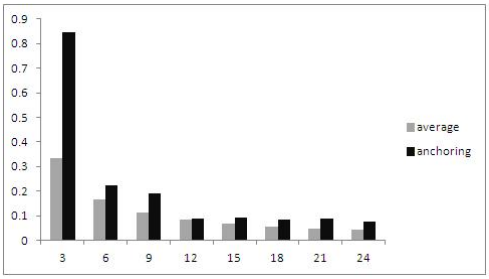


Figure 5.7: Anchoring effect as a function of the group size. Reliabilities coming from a β-distribution. First speaker 0.5.

5.7 the left bar shows (for each group size) the probability that the final result is expected to be closest to that of the first speaker (average) and the right bar shows the result of simulations.

As the simulation results suggest, the anchoring effect happens for the inhomogeneous groups as well as the homogeneous ones. Here again the incremental

process of updating will result in the first speaker receiving the highest weight in the formation of the group's consensus. Of course the anchoring effect is much more apparent in the homogeneous groups and as the inhomogeneity of the group increases the anchoring effect decreases. In particular the effect will become less evident if we move to a deliberation model with a more fine graded assessment of reliabilities that is a model with a higher number of reliability brackets.

5.4 Conclusion

The model we propose here is a simple incremental updating procedure which seems close (for our purpose) to how the deliberations proceed. It is inspired by the Lehrer-Wagner model and presents a way how to calculate the Lehrer Wagner matrix. To wit, after the first round, a Lehrer Wagner matrix is generated and it is easy to see that it satisfies the normalisation condition. The results suggest that the such updating procedures will indeed result in the emergence of anchoring effect even for fully rational agents with no cognitive limitations.

Since anchoring effect is in most cases an undesired bias of the deliberation process, it would be relevant to investigate ways to prevent it which we hope to do in future work. There are a couple immediate strategies that come to mind in this regard.

- One option is to decide randomly who will speak first, second, and third etc. This does not however prevent the anchoring effect, but it will be avoided that a specific group member can get the chance to affect the final result of the deliberation by strategically speaking first.
- A second option is to give those who speak earlier a lower weight, so that their initial assignment does not have too much of an impact. One way to model this is to assign weights which are inversely proportional to the step in the corresponding round. A justified process to do so without introducing new unpleasant characteristics however, is by no means obvious.

5.5 Appendix

5.5.1 Proof of Theorem 5.3.2

In this appendix we give the proof of the Theorem 5.3.2.

Lemma 5.5.1 *For* $1 \leq m \leq n$,

$$B_n^m . B_n^{m-1} \ldots . B_n^1 = \begin{pmatrix} \bar{B}^m & \mathbf{0} \\ C^m & I_{n-m} \end{pmatrix},$$

where

$$\bar{B}^m = \begin{pmatrix} 2^{m-1}+1 & 2^{m-2} & ... & 2^0 \\ 2^{m-1} & 2^{m-2}+1 & ... & 2^0 \\ 2^{m-1} & 2^{m-2} & ... & 2^0 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 2^{m-1} & 2^{m-2} & ... & 2^0+1 \end{pmatrix}_{m\times m}$$

$\mathbf{0}$ *is an* $m \times (n-m)$ *zero matrix,* I_{n-m} *is the unite* $(n-m)\times(n-m)$ *matrix and*

$$C^m = \begin{pmatrix} 2^{m-1} & 2^{m-2} & ... & 2^0 \\ 2^{m-1} & 2^{m-2} & ... & 2^0 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 2^{m-1} & 2^{m-2} & ... & 2^0 \end{pmatrix}_{(n-m)\times m}$$

Proof For $m = 1$ the result is trivially true. Suppose it is true for m and we shall show that it holds for $m+1$.

$$B_n^{m+1}.B_n^{m-1}.....B_n^1 = B_n^{m+1}.\begin{pmatrix} \bar{B}^m & \mathbf{0} \\ C^m & I_{n-m} \end{pmatrix}$$

$$= (A_n^{m+1} + I_n).\begin{pmatrix} \bar{B}^m & \mathbf{0} \\ C^m & I_{n-m} \end{pmatrix}$$

$$= A_n^{m+1}.\begin{pmatrix} \bar{B}^m & \mathbf{0} \\ C^m & I_{n-m} \end{pmatrix} + I_n.\begin{pmatrix} \bar{B}^m & \mathbf{0} \\ C^m & I_{n-m} \end{pmatrix} =$$

$$\begin{pmatrix} 2^{m-1} & 2^{m-2} & ... & 2^0 & 1 & 0 & ... & 0 \\ 2^{m-1} & 2^{m-2} & ... & 2^0 & 1 & 0 & ... & 0 \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ 2^{m-1} & 2^{m-2} & ... & 2^0 & 1 & 0 & ... & 0 \\ 2^{m-1} & 2^{m-2} & ... & 2^0 & 1 & 0 & ... & 0 \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ 2^{m-1} & 2^{m-2} & ... & 2^0 & 1 & 0 & ... & 0 \end{pmatrix} + \begin{pmatrix} 2^{m-1}+1 & 2^{m-2} & ... & 2^0 & 0 & ... & 0 \\ 2^{m-1} & 2^{m-2}+1 & ... & 2^0 & 0 & ... & 0 \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 2^{m-1} & 2^{m-2} & ... & 2^0+1 & 0 & ... & 0 \\ 2^{m-1} & 2^{m-2} & ... & 2^0 & 1 & ... & 0 \\ . & . & . & . & . & ... & . \\ . & . & . & . & . & ... & . \\ 2^{m-1} & 2^{m-2} & ... & 2^0 & 0 & ... & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 2^m+1 & 2^{m-1} & ... & 2^1 & 1 & 0 & ... & 0 \\ 2^m & 2^{m-1}+1 & ... & 2^1 & 1 & 0 & ... & 0 \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ 2^m & 2^{m-1} & ... & 2^1+1 & 1 & 0 & ... & 0 \\ 2^m & 2^{m-1} & ... & 2^1 & 2 & 0 & . & 0 \\ . & . & . & . & 1 & 1 & . & 0 \\ . & . & . & . & 1 & 0 & . & 0 \\ 2^m & 2^{m-1} & ... & 2^1 & 1 & 0 & ... & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 2^m+1 & 2^{m-1} & ... & 2^1 & 2^0 & 0 & ... & 0 \\ 2^m & 2^{m-1}+1 & ... & 2^1 & 2^0 & 0 & ... & 0 \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ 2^m & 2^{m-1} & ... & 2^1+1 & 2^0 & 0 & ... & 0 \\ 2^m & 2^{m-1} & ... & 2^1 & 2^0+1 & 0 & . & 0 \\ . & . & . & . & 1 & 1 & . & 0 \\ . & . & . & . & 1 & 0 & . & 0 \\ 2^m & 2^{m-1} & ... & 2^1 & 2^0 & 0 & ... & 1 \end{pmatrix} = \begin{pmatrix} \bar{B}^{m+1} & \mathbf{0} \\ C^{m+1} & I_{n-(m+1)} \end{pmatrix}.$$

this completes the proof of Lemma 5.5.1. □

Corollary 1 *The result of updating the assignments $\vec{V}^{(0)}$ through one round of deliberation is given by*

$$\vec{V}^{(1)} = \frac{1}{2^n}\left(B_n^n.B_n^{n-1}....,B_n^1\right)\vec{V}^{(0)} =$$

$$\left(\frac{1}{2^n}\begin{pmatrix} 2^{n-1} & 2^{n-2} & ... & 2^0 \\ 2^{n-1} & 2^{m-2} & ... & 2^0 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 2^{n-1} & 2^{n-2} & ... & 2^0 \end{pmatrix} + \frac{1}{2^n}I_n\right)\vec{V}^{(0)}.$$

Proposition 5.5.2 *Let*

$$B = \frac{1}{2^n}\prod_{i=1}^{n} B_n^i - \frac{1}{2^n}I_n,$$

then, the result of updating the assignments $\vec{V}^{(0)}$ *through* k *round of deliberation is given by*

$$\vec{V}^{(k)} = \left(\frac{1}{2^n}\left(B_n^n.B_n^{n-1}\ldots,B_n^1\right)\right)^k \vec{V}^{(0)} = \left(\sum_{t=1}^{k}\binom{k}{t}\frac{(2^n-1)^{t-1}}{(2^n)^{k-1}}B + \frac{1}{2^{nk}}I_n\right).\vec{V}^{(0)}$$

Proof Let $b_i = \sum_{j=1}^{n} < B >_{ij}$. Notice that by Lemma 5.5.1, in matrix B all rows are equal so $b_1 = b_2 = \ldots = b_n = b$ where $b = \sum_{i=1}^{n} 2^{-i} = 1 - \frac{1}{2^n}$, moreover we have

$$B^k = b^{k-1}B.$$

$$\vec{V}^{(k)} = \left(\frac{1}{2^n}\prod_{i=1}^{n} B_n^i\right)^k \vec{V}^{(0)} =$$

$$= \left(B + \frac{1}{2^n}I_n\right)^k \vec{V}^{(0)} = \left(\frac{1}{2^{nk}}I_n + \sum_{t=1}^{k}\binom{k}{t}(\frac{1}{2^n}I_n)^{k-t}B^t\right)\vec{V}^{(0)}$$

$$= \left(\frac{1}{2^{nk}}I_n + \sum_{t=1}^{k}\binom{k}{t}(\frac{1}{2^n}I_n)^{k-t}(b^{t-1}B)\right)\vec{V}^{(0)} = \left(\frac{1}{2^{nk}}I_n + \sum_{t=1}^{k}\binom{k}{t}\frac{(1-\frac{1}{2^n})^{t-1}}{2^{n(k-t)}}B\right)\vec{V}^{(0)}$$

$$= \left(\frac{1}{2^{nk}}I_n + \sum_{t=1}^{k}\binom{k}{t}\frac{(2^{n-1})^{t-1}}{2^{n(k-1)}}B\right)\vec{V}^{(0)}$$

as required. □

Proof Theorem 5.3.2
Let $\vec{V}$ be asymptotic result of deliberation and w_i be the weight assigned to the i^{th} speaker in the limit. By Proposition 5.5.2,

$$\vec{V} = \lim_{k\to\infty}\left(\sum_{t=1}^{k}\binom{k}{t}\frac{(2^n-1)^{t-1}}{(2^n)^{k-1}}B + \frac{1}{2^{nk}}I_n\right).\vec{V}^{(0)} = \lim_{k\to\infty}\left(\sum_{t=1}^{k}\binom{k}{t}\frac{(2^n-1)^{t-1}}{(2^n)^{k-1}}B\right).\vec{V}^{(0)}$$

$$= dB.V^{(0)}$$

where $d = \lim_{k\to\infty}\sum_{t=1}^{k}\binom{k}{t}\frac{(2^n-1)^{t-1}}{(2^n)^{k-1}}$, then by Lemma 5.5.1 the weight assigned to the i^{th} speaker is given by $w_i = d.2^{-i}$. Thus we have

$$w_1 \geq w_2 \geq \ldots \geq w_n.$$

□

5.5.2 Stability Result

Next we will investigate the stability of our results for the homogeneous groups by showing that the results still hold for small changes in the equal weight assumption. Here the group members consider small deviation from the equal weights by assigning $\frac{1-\epsilon}{2}$ to the speaker and $\frac{1+\epsilon}{2}$ to themselves. As before, after the i^{th} group member speaks, everyone will update their assignment as a weighted average of their current value and that announced by the i^{th} speaker while assigning a slightly higher/lower weight to themselves as opposed to the speaker. This process can again be represented by matrix multiplication as before by replacing the matrices B_n^i with

$$B_n^{i\prime} = B_n^i - \epsilon E_n^i$$

where

$$E_n^j = B_n^j - 2A_n^j.$$

Thus for example, in group of size 3, the matrix

$$B_3^2 = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 1 & 1 \end{pmatrix}$$

will be replaced by

$$B_3^{2\prime} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 1 & 1 \end{pmatrix} + \epsilon \left(\begin{pmatrix} 1 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 1 & 1 \end{pmatrix} - 2 \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix} \right)$$

$$= \begin{pmatrix} 1 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 1 & 1 \end{pmatrix} + \epsilon \begin{pmatrix} 1 & -1 & 0 \\ 0 & 0 & 0 \\ 0 & -1 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1+\epsilon & 1-\epsilon & 0 \\ 0 & 2 & 0 \\ 0 & 1-\epsilon & 1+\epsilon \end{pmatrix}.$$

In this new setting we shall again have that the result of one round of deliberation is given as

$$\vec{V}^{(1)} = \frac{1}{2^n} \prod_{i=1}^{n} B_n^{i\prime}.\vec{V}^{(0)}$$

and after k round of deliberation we shall have

$$\vec{V}^{(k)} = \left(\frac{1}{2^n} \prod_{i=1}^{n} B_n^{i\prime} \right)^k \vec{V}^{(0)}.$$

We shall first prove the corresponding versions of Lemma 5.5.1 for this setting.

Lemma 5.5.3 *Let A_n^i's be the matrices with 1 on the i^{th} column and zero elsewhere and $B_n^j = I_n + A_n^i$ as before. Then*

$$\prod_{k=j+1}^{n} B_n^k A_n^j \prod_{k=1}^{j-1} B_n^k = 2^{n-j} A_n^j + \sum_{k=1}^{j-1} 2^{n-1-k} A_n^k.$$

Proof

$$\prod_{k=j+1}^{n} B_n^k A_n^j \prod_{k=1}^{j-1} B_n^k = \prod_{k=j+1}^{n} (I_n + A_n^k) A_n^j \prod_{k=1}^{j-1} B_n^k.$$

First notice that for all $i, j = 1, \ldots, n$ we have $A_n^i . A_n^j = A_n^j$, thus

$$\prod_{k=j+1}^{n} B_n^k A_n^j \prod_{k=1}^{j-1} B_n^k = \prod_{k=j+1}^{n} (I_n + A_n^k) A_n^j \prod_{k=1}^{j-1} B_n^k = 2^{n-j} A_n^j \prod_{k=1}^{j-1} B_n^k.$$

By Lemma 5.5.1 we have

$$\prod_{k=1}^{j-1} B_n^k = I_n + \sum_{k=1}^{j-1} 2^{j-1-k} A_n^k$$

thus

$$\prod_{k=j+1}^{n} B_n^k A_n^j \prod_{k=1}^{j-1} B_n^k = \prod_{k=j+1}^{n} (I_n + A_n^k) A_n^j \prod_{k=1}^{j-1} B_n^k$$

$$= 2^{n-j} A_n^j \left(I_n + \sum_{k=1}^{j-1} 2^{j-1-k} A_n^k \right)$$

$$= 2^{n-j} A_n^j + 2^{n-j} A_n^j \sum_{k=1}^{j-1} 2^{j-1-k} A_n^k$$

$$= 2^{n-j} A_n^j + \sum_{k=1}^{j-1} 2^{n-1-k} A_n^k$$

as required. □

Next we have a modified version of the Theorem 5.5.2:

Proposition 5.5.4

$$\frac{1}{2^n} \prod_{i=1}^{n} B_n^{i\prime} = \frac{1}{2^n} \prod_{i=1}^{n} B_n^i + \frac{n\epsilon}{2^n} I_n + \epsilon \left(\sum_{j=1}^{n} \frac{(j-2)}{2^j} A_n^j \right) + \frac{O(\epsilon^2)}{2^n}$$

Proof $\frac{1}{2^n}\prod_{i=1}^{n} B_n^{i\prime} = \frac{1}{2^n}\prod_{i=1}^{n}(B_n^i + \epsilon E_n^i)$

$$= \frac{1}{2^n}\left(\prod_{i=1}^{n} B_n^i + \epsilon\sum_{j=1}^{n}\left(\prod_{k=j+1}^{n} B_n^k\right)E_n^j\left(\prod_{k=1}^{j-1} B_n^k\right) + O(\epsilon^2)\right)$$

$$= \frac{1}{2^n}\left(\prod_{i=1}^{n} B_n^i + \epsilon\sum_{j=1}^{n}\left(\prod_{k=j+1}^{n} B_n^k\right)(B_n^j - 2A_n^j)\left(\prod_{k=1}^{j-1} B_n^k\right) + O(\epsilon^2)\right)$$

$$= \frac{1}{2^n}\prod_{i=1}^{n} B_n^i + \frac{\epsilon}{2^n}\sum_{j=1}^{n}\left(\prod_{k=1}^{n} B_n^k - 2\prod_{k=j+1}^{n} B_n^k A_n^j \prod_{k=1}^{j-1} B_n^k\right) + \frac{O(\epsilon^2)}{2^n}$$

$$= \frac{1}{2^n}\prod_{i=1}^{n} B_n^i + \frac{\epsilon}{2^n}\sum_{j=1}^{n}\left(\prod_{k=1}^{n} B_n^k - 2\left(2^{n-j}A_n^j + \sum_{k=1}^{j-1} 2^{n-1-k}A_n^k\right)\right) + \frac{O(\epsilon^2)}{2^n}$$

$$= \frac{1}{2^n}\prod_{i=1}^{n} B_n^i + \frac{\epsilon}{2^n}\sum_{j=1}^{n}\left(\left(I_n + \sum_{k=1}^{n} 2^{n-k}A_n^k\right) - \left(2^{n-j+1}A_n^j + \sum_{k=1}^{j-1} 2^{n-k}A_n^k\right)\right) + \frac{O(\epsilon^2)}{2^n}$$

$$= \frac{1}{2^n}\prod_{i=1}^{n} B_n^i + \frac{\epsilon}{2^n}\sum_{j=1}^{n}\left(I_n + \sum_{k=j}^{n} 2^{n-k}A_n^k - 2^{n-j+1}A_n^j\right) + \frac{O(\epsilon^2)}{2^n}$$

$$= \frac{1}{2^n}\prod_{i=1}^{n} B_n^i + \frac{n\epsilon}{2^n}I_n + \frac{\epsilon}{2^n}\sum_{j=1}^{n}\left(\sum_{k=j}^{n} 2^{n-k}A_n^k - 2^{n-j+1}A_n^j\right) + \frac{O(\epsilon^2)}{2^n}$$

$$= \frac{1}{2^n}\prod_{i=1}^{n} B_n^i + \frac{n\epsilon}{2^n}I_n + \frac{\epsilon}{2^n}\sum_{j=1}^{n}\left(\sum_{k=j+1}^{n} 2^{n-k}A_n^k - 2^{n-j}A_n^j\right) + \frac{O(\epsilon^2)}{2^n}$$

$$= \frac{1}{2^n}\prod_{i=1}^{n} B_n^i + \frac{n\epsilon}{2^n}I_n + \frac{\epsilon}{2^n}\left(\sum_{j=1}^{n}\left(\sum_{k=j+1}^{n} 2^{n-k}A_n^k\right) - \left(\sum_{j=1}^{n} 2^{n-j}A_n^j\right)\right) + \frac{O(\epsilon^2)}{2^n}$$

$$= \frac{1}{2^n}\prod_{i=1}^{n} B_n^i + \frac{n\epsilon}{2^n}I_n + \frac{\epsilon}{2^n}\left(\sum_{j=1}^{n}(j-1)2^{n-j}A_n^j - \sum_{j=1}^{n} 2^{n-j}A_n^j\right) + \frac{O(\epsilon^2)}{2^n}$$

$$= \frac{1}{2^n}\prod_{i=1}^{n} B_n^i + \frac{n\epsilon}{2^n}I_n + \frac{\epsilon}{2^n}\left(\sum_{j=1}^{n}(j-2)2^{n-j}A_n^j\right) + \frac{O(\epsilon^2)}{2^n}$$

$$= \frac{1}{2^n}\prod_{i=1}^{n} B_n^i + \frac{n\epsilon}{2^n}I_n + \epsilon\left(\sum_{j=1}^{n}\frac{(j-2)}{2^j}A_n^j\right) + \frac{O(\epsilon^2)}{2^n}$$

$$= \frac{1}{2^n}\prod_{i=1}^{n} B_n^i + \epsilon \begin{pmatrix} \frac{-1}{2}+\frac{n}{2^n} & 0 & \frac{1}{2^3} & \frac{2}{2^4} & \cdots & \frac{n-2}{2^n} \\ \frac{-1}{2} & 0+\frac{n}{2^n} & \frac{1}{2^3} & \frac{2}{2^4} & \cdots & \frac{n-2}{2^n} \\ \frac{-1}{2} & 0 & \frac{1}{2^3}+\frac{n}{2^n} & \frac{2}{2^4} & \cdots & \frac{n-2}{2^n} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \frac{-1}{2} & 0 & \frac{1}{2^3} & \frac{2}{2^4} & \cdots & \frac{n-2}{2^n}+\frac{n}{2^n} \end{pmatrix} + \frac{O(\epsilon^2)}{2^n}$$

where the fifth equality is given by Lemma 5.5.3. □

We are now in a position to prove the equivalent of Theorem 5.3.2 for the agents that can make small deviation from the equal weight assumptions.

Theorem 5.5.1 *For ϵ sufficiently small, the process of deliberation described above with weights set as $\frac{1-\epsilon}{2}$ and $\frac{1+\epsilon}{2}$ will end in consensus. That is if*

$$\vec{V} = \lim_{k\to\infty}\left(\frac{1}{2^n}\prod_{i=1}^{n} B_n^{i\prime}\right)^k \vec{V}^{(0)}$$

then $<\vec{V}>_i = <\vec{V}>_j$ for $i,j = 1,\ldots,n$. Moreover in the final consensus the opinion of the first speaker receives the highest weight compared to other group members.

Proof Notice that $\frac{1}{2^n}\prod_{i=1}^{n} B_n^{i\prime}$ is essentially a weight matrix in the sense of Lehrer-Wagner model and thus the convergence in the limit follows for the same reason as in the Lehrer-Wagner model. We will now show that the first speaker receives the highest weight in the final consensus. By Proposition 5.5.4 we have

$$\vec{V} = \lim_{k\to\infty}\left(\frac{1}{2^n}\prod_{i=1}^{n} B_n^{i\prime}\right)^k \vec{V}^{(0)}$$

$$= \lim_{k\to\infty}\left(\frac{1}{2^n}\prod_{i=1}^{n} B_n^{i} + \frac{n\epsilon}{2^n} I_n + \epsilon\left(\sum_{j=1}^{n}\frac{(j-2)}{2^j}A_n^j\right) + \frac{O(\epsilon^2)}{2^n}\right)^k V^{(0)}$$

$$= \lim_{k\to\infty}\left(\sum_{n_1+n_2+n_3+n_4=k}\left(\frac{1}{2^n}\prod_{i=1}^{n} B_n^{i}\right)^{n_1}\left(\frac{n\epsilon}{2^n}I_n\right)^{n_2}\left(\epsilon\sum_{j=1}^{n}\frac{(j-2)}{2^j}A_n^j\right)^{n_3}\left(\frac{O(\epsilon^2)}{2^n}\right)^{n_4}\right)V^{(0)}.$$

Any term with $n_2 \geq 2$ or $n_3 \geq 2$ or $n_4 \geq 1$ include a term in the order of $O(\epsilon^2)$ and thus can be ignored, the same is true for the term with $n_2 = n_3 = 1$ so

$$\vec{V} = \lim_{k\to\infty}\left(\frac{1}{2^n}\prod_{i=1}^{n} B_n^{i\prime}\right)^k \vec{V}^{(0)}$$

$$= \lim_{k\to\infty}\left(\left(\frac{1}{2^n}\prod_{i=1}^{n}B_n^i\right)^k + \left(\frac{1}{2^n}\prod_{i=1}^{n}B_n^i\right)^{k-1}\left(\frac{n\epsilon}{2^n}I_n\right) + \left(\frac{1}{2^n}\prod_{i=1}^{n}B_n^i\right)^{k-1}\left(\epsilon\sum_{j=1}^{n}\frac{(j-2)}{2^j}A_n^j\right)\right)V^{(0)}$$

$$= \lim_{k\to\infty}\left(\frac{1}{2^n}\prod_{i=1}^{n}B_n^i\right)^k + \lim_{k\to\infty}\left(\frac{1}{2^n}\prod_{i=1}^{n}B_n^i\right)^{k-1}\left(\frac{n\epsilon}{2^n}I_n + \epsilon\sum_{j=1}^{n}\frac{(j-2)}{2^j}A_n^j\right)$$

$$= d\,B + d\,B\left(\frac{n\epsilon}{2^n}I_n + \epsilon\sum_{j=1}^{n}\frac{(j-2)}{2^j}A_n^j\right) = d\,B\left((1+\frac{n\epsilon}{2^n})I_n + \epsilon\sum_{j=1}^{n}\frac{(j-2)}{2^j}A_n^j\right)$$

$$= d(1+\frac{n\epsilon}{2^n})B + d\,\epsilon(B.\sum_{j=1}^{n}\frac{(j-2)}{2^j}A_n^j)$$

where $d = \lim_{k\to\infty}\sum_{t=1}^{k}\binom{k}{t}\frac{(2^n-1)^{t-1}}{(2^n)^{k-1}}$ and $B = \frac{1}{2^n}\prod_{i=1}^{n}B_n^i - \frac{1}{2^n}I_n$ as in Theorem 5.3.2. Notice that B and $\sum_{j=1}^{n}\frac{(j-2)}{2^j}A_n^j$ are both matrices with all rows equal thus

$$B.\sum_{j=1}^{n}\frac{(j-2)}{2^j}A_n^j = (\sum_{k=1}^{n}B_{ik})(\sum_{j=1}^{n}\frac{(j-2)}{2^j}A_n^j) = (1-\frac{1}{2^n})(\sum_{j=1}^{n}\frac{(j-2)}{2^j}A_n^j)$$

where the last equality is derived from Corollary 1. Thus

$$\vec{V} = \lim_{k\to\infty}\left(\frac{1}{2^n}\prod_{i=1}^{n}B_n^{i\prime}\right)^k\vec{V}^{(0)} = \left(d(1+\frac{n\epsilon}{2^n})B + d\epsilon(B.\sum_{j=1}^{n}\frac{(j-2)}{2^j}A_n^j)\right)V^{(0)}$$

$$= d\left((1+\frac{n\epsilon}{2^n})B + \epsilon(1-\frac{1}{2^n})(\sum_{j=1}^{n}\frac{(j-2)}{2^j}A_n^j)\right)V^{(0)}$$

The weight assigned to the i^{th} member will be the sum of the weights assigned through B and $\sum_{j=1}^{n}\frac{(j-2)}{2^j}A_n^j)$ that is

$$d\,(1+\frac{n\epsilon}{2^n})\,2^{-i} + d\,\epsilon\,(1-\frac{1}{2^n})\frac{(i-2)}{2^i}$$

For n large enough we have $\frac{n\epsilon}{2^n} \approx 0$ and hence w_i (the weight assigned to i^{th} speaker) will be

$$d.2^{-i} + d\epsilon(1-\frac{1}{2^n})\frac{(i-2)}{2^i}.$$

□

It is now easy to check that for $\epsilon \le 1/2$ we have $w_1 \ge w_i$ for $i = 1,\ldots,n$; that means the first speaker receives the highest weight.

Chapter 6

Conclusion

In this thesis, we studied four problems of long standing interest in philosophy using mathematical and computational methods. The problems we investigated here belong to different philosophical discipline and each has attracted attention from different parts of philosophical community. The treatment of inconsistencies has been of interest to both philosophers and mathematicians. Logic of conditionals has been a main stream topic in philosophy for at least several decades and the study of rational deliberation and related issues has been of interests to epistemologists, political philosophers and philosophers of social sciences.

Our goal was to demonstrate that a wide range of problems in philosophy can be fruitfully studied in the approach that has come to be known as the scientific philosophy. That is by means of scientific methodology such as application of mathematical modelling and computational methods. The goal was to show the advantages of such methodology in addressing, or at least further clarifying, many issues and concerns with respect to the philosophical investigation of these topics. From these, we hope that the advantages of the scientific approach to philosophical problems of similar nature would be evident.

As already pointed out, our advocacy of the the scientific methods is by no means intended to imply that all problems of philosophical nature should or even can be adequately represented in a formal machinery or investigated by means of formal and computational methods. The aim was to further motivate the benefits of employing mathematical and computational methods along side the traditional philosophical toolbox as they allow for precise and exact argumentations and extensive study of relevant solution spaces.

The studies presented in this thesis, each introduce further work in their respective areas and the methodologies employed for the study of these issues present immediate generalisations and seem to facilitate the investigation of more

general cases. The formal machinery used for the study of probabilistic consequence relation and modelling the learning of indicative conditionals seem to provide natural extensions to be considered for the more general cases. The models developed for rational deliberation and the study of the anchoring effect can also be considered as base models that can be further developed into more complex ones by introducing more relevant factors.

More precisely, the work presented on dealing with inconsistent evidence begs for further investigation of distance measures and better justified updating mechanisms. In particular, we hope to study mechanisms for updating the weights of information when dealing with prioritised evidence. That is, how to update the degree of entrenchment of the information in the belief set (along with their probabilities).

The work on leaning indicative conditionals should be further developed to better understand the effect of introducing causal structures and to investigate the similar approach for dealing with counter-factual cases. We also hope to investigate whether KL distance can be replaced with another distance measure with better mathematical properties including, for example, symmetry.

The models developed for group deliberation and the study of the anchoring effect in deliberations are toy models that include many abstractions. We hope to further develop these models to incorporate a wider range of relevant characteristics of the deliberating agents and to allow for private communications by the agents both amongst themselves as well as with sources outside the group.

Bibliography

[van Aken et al.2004] van Aaken, A., List, C. and Luetge, C. (eds) (2004). Deliberation and Decision: Economics, Constitutional Theory and Deliberative Democracy. Ashgate.

[Alchourron et. al.1985] Alchourrn, C. E., Gadenfors, P., and Makinson, D. (1985). "On the logic of theory change: Partial meet contraction and revision functions", *Journal of Symbolic Logic* 50: 510-530.

[3] Angere, S. (2010). Knowledge in a Social Network. Preprint. Department of Philosphy. Lund University.

[Arlo-Costa1990] Arlo-Costa, H. (1990). "Conditionals and Monotonic Belief Revisions: The Success Postulate?", *Studia Logica*, 49(4): 557–566.

[Bohman & Rehg 1997] Bohman, J. and Rehg, W. (eds) (1997). Deliberative Democracy. *s on Reason and Politics*. Cambridge (Mass.): MIT Press.

[Bovens & Hartmann 2003] Bovens, L. and Hartmann, S. (2003). *Bayesian Epistemology*. Oxford: Oxford University Press.

[Cohen 1986] Cohen, J. (1986). An Epistemic Conception of Democracy. *Ethics* 97(1): 26–38. The University of Chicago Press.

[Cohen 1989a] Cohen, J. (1989). Deliberation and Democratic Legitimacy. *The Good Polity: Normative Analysis of the State* 17–34 Hamlin A and Pettit P. (eds). Oxford: Blackwell.

[Cohen 1989b] Cohen, J. (1989). The Economic Basis of Deliberative Democracy. *Social Philosophy and Policy* 6(2): 25–50.

[da Costa 1974] da Costa, N.C.A. (1974). On the Theory of Inconsistent Formal Systems, *Notre Dame Journal of Formal Logic* 15(4): 497-510.

[da Costa 1989] da Costa, N.C.A. and Subrahmanian, V.S. (1989). "Paraconsistent logic as a formalism for reasoning about inconsistent knowledge bases", *Artificial Intelligence in Medicine* 1: 167-174.

[da Costa 1998] da Costa, N.C.A. (1998), "Paraconsistent logic", *Stanisaw Jakowski Memorial Symposium* 29–35.

[Christiano 1997] Christiano, T. (1997). The Significance of Public Deliberation. *Bohman and Rehg (1997)*: 243–278.

[Christiano 2004] Christiano, T. (2004). The Authority of Democracy. *Journal of Political Philosophy* 12(3): 266–290.

[Diaconis & Zabell 1982] Diaconis, P. and S. Zabell (1982). Updating Subjective Probability, *Journal of the American Statistical Association* 77: 822–830.

[Dietrich & Spiekermann 2010] Dietrich, F. and Spiekermann, K. (2010). Epistemic democracy with defensible premises. Available at http://www.franzdietrich.net/Papers/DietrichSpiekermann-EpistemicDemocracy.pdf.

[Dietrich 2006] Dietrich, F. (2006). General Representation of Epistemically Optimal Procedures. *Social Choice and Welfare* 26(2): 263–283.

[Douven 2012] Douven, I. (2012). Learning Conditional Information, *Mind & Language* 27 (3): 239–263.

[Douven & Dietz 2011] Douven, I. and R. Dietz (2011). A Puzzle about Stalnaker's Hypothesis, *Topoi* 30: 31–37.

[Douven & Romeijn 2012] Douven, I. and J. W. Romeijn (2012). A New Resolution of the Judy Benjamin Problem, *Mind* 120 (479): 637–670.

[Dryzek 1990] Dryzek, J. (1990). *Discursive Democracy*. Cambridge: Cambridge University Press.

[Dryzek & Niemeyer 2010] Dryzek, J. and S. Niemeyer (2010). *Foundations and Frontiers of Deliberative Governance*. Oxford: Oxford University Press.

[Duran & Arnold 2013] Duran, J. M. and Arnold, E. (2013). *Computer Simulations and the Changing Face of Scientific Experimentation*. Cambridge Scholars Publishing.

[Elga 2007] Elga, A. (2007). Reflection and Disagreement. *Noûs* 41(3): 478–502. Also in: P. Grim, I. Flora, and A. Plakias (eds.): *The Philosopher's Annual* 27 (2007).

[Elster 1998] Elster, J. (eds) (1998). *Deliberative Democracy* Cambridge: Cambridge University Press.

[Estlund 1993] Estlund, D. (1993). Whos Afraid of Deliberative Democracy? On the Strategic/Deliberative Dichotomy in Recent Constitutional Jurisprudence. *Texas Law Review* 71:1437-1477.

[Estlund 1994] Estlund, D. (1994). Opinion Leaders, Independence, and Condorcets Jury Theorem. *Theory and Decision* 36(2):131–162.

[Estlund 1997] Estlund, D. (1997). Beyond Fairness and Deliberation: The Epistemic Dimension of Democratic Authority. *Bohman and Rehg (1997)*: 173–204.

[Estlund 2009] Estlund, D. (2009). *Democratic Authority: A Philosophical Framework.* Princeton: Princeton University Press.

[Falkenhainer et. al. 1989] Falkenhainer, B., Forbus, K. D., and Gentner, D. (1989). "The Structure mapping Engine: Algorithms and Examples." *Artificial Intelligence*, 41:1–63.

[Fearon 1998] Fearon, J. D. (1998). Deliberation as Discussion. *Deliberative Democracy*:44–68. Elster (eds) Cambridge: Cambridge University Press.

[Fitelson & Zalta 2007] Fitelson, B. and Zalta, E. N. (2007). "Steps Toward a Computational Metaphysics", *Journal of Philosophical Logic*, 36(2): 227–247.

[van Fraassen 1981] Fraassen, B. van (1981). A Problem for Relative Information Minimizers in Probability Kinematics, *British Journal for the Philosophy of Science* 32: 375–379.

[Goodin 2008] Goodin, R.E. (2008). *Innovating Democracy: Democratic Theory And Practice After The Deliberative Turn.* Oxford: Oxford University Press.

[Hartmann et. al. 2009] Hartmann, S., C. Martini and J. Sprenger (2009). Consensual Decision-Making Among Epistemic Peers. *Episteme* 6: 110–129.

[Hegselmann & Krause 2002] Hegselmann, R. and U. Krause (2002). Opinion Dynamics and Bounded Confidence: Models, Analysis and Simulation. *Journal of Artificial Societies and Social Simulation* 5(3).

[Hegselmann & Krause 2006] Hegselmann, R. and U. Krause (2006). Truth and Cognitive Division of Labour: First Steps Towards a Computer Aided Social Epistemology. *Journal of Artificial Societies and Social Simulation* 9(3).

[Hegselmann & Krause 2009] Hegselmann, R. and U. Krause (2009). Deliberative Exchange, Truth, and Cognitive Division of Labour: A Low-Resolution Modeling Approach. *Episteme* 6: 130–144.

[Holyoak & Thagard 1989] Holyoak, K. J., and Thagard, P. (1989). "Analogical Mapping by Constraint Satisfaction." *Cognitive Science*, 13: 295–355.

[Holyoak & Thagard 1995] Holyoak, K. J., and Thagard, P. (1995). *Mental Leaps: Analogy in Creative Thought*. Cambridge, MA: MIT Press/Bradford Books.

[Kahneman et. al. 2006] Kahneman, D., Krueger, A., Schkade, D., Schwarz, N., and Stone, A. (2006). Would you be Happier if you Were Richer? A Focusing Illusion. *Science* 312(5782): 1908–10.

[Kern-Isberner 2001] Kern-Isberner, G. (2001). *Conditionals in Nonmonotonic Reasoning and Belief Revision.* Berlin: Springer.

[Knight 2002] Knight, K. M. (2002). Doctoral Thesis, University of Manchester, UK. [see http://www.maths.manchester.ac.uk/ jeff]

[Kulkarni & Simon 1988] Kulkarni, D. and Simon, H. (1988). "The Processes of Scientific Discovery: The Strategy of Experimentation." *Cognitive Science*, 12:139–175.

[Ladyman & Ross 2009] Ladyman, J., and Ross, D. (2009). *Every Thing Must Go: Metaphysics Naturalized*. Oxford University Press.

[Langley et. al. 1987] Langley, P., Simon, H., Bradshaw, G., and Zytkow, J. (1987). *Scientific Discovery*. Cambridge, MA: MIT Press/Bradford Books.

[Lehrer 1976] Lehrer, K. (1976). When Rational Disagreement is Impossible. *Noûs* 10: 327–332.

[Lehrer & Wagner 1981] Lehrer, K. and C. Wagner (1981). *Rational Consensus in Science and Society*. Dordrecht: Reidel.

[Leitgeb 2013] Leitgeb, H., (2013). "Scientific Philosophy, Mathematical Philosophy, and All That", *Metaphilosophy*, 44(3): 267–275.

[Lewis 1976] Lewis, D. (1976). Probabilities of Conditionals and Conditional Probabilities, *Philosophical Review* 85: 297–315.

[List & Goodin 2001] List, C. and Goodin, R.E. (2001). Epistemic democracy: Generalizing the Condorcet Jury Theorem. *Journal of Political Philosophy* 9: 277–306.

[Maddy 2009] Maddy, P. (2009). *Second Philosophy*, Oxford University Press.

[Manin 1987] Manin B. (1987). On Legitimacy and Political Deliberation. *Political Theory* 15(3): 338–368.

[Marti 2006] Marti, J.L. (2006). The Epistemic Conception of Deliberative Democracy Defended. Reasons, Rightness and Equal Political Liberty. *Deliberative Democracy and Its Discontents. National and Post-national Challenges*. Besson, S. and Marti, J.L. (eds). London: Ashgate.

[Nino 1996] Nino C.S. (1996). *The Constitution of Deliberative Democracy*. New Haven: Yale University Press.

[Olsson 2011] Olsson, E.J. (2011). A Bayesian simulation model of group deliberation and polarization. *Bayesian Argumentation*. Zenker, F. (ed.). Synthese Library, Springer Verlag.

[57] Olsson, E.J. (2011b). A Simulation Approach to Veritistic Social Epistemology. *Episteme* 8(2): 127–143.

[58] Vallinder, A. and E.J. Olsson (2013). Do Computer Simulations Support the Argument from Disagreement? *Synthese* 190(8): 1437–1454.

[Paris & Vencovska 1989] Paris, J.B. and Vencovska, A. (1989). "Maximum Entropy And Inductive Inference", J. Skilling (ed.) *Maximum Entropy and Bayesian Methods* 397–403.

[Paris 2004] Paris, J.B. (2004). "Deriving information from inconsistent knowledge bases: A completeness theorem", *Logic Journal of the IGPL* 12: 345–353.

[Paris et. al. 2008] Paris, J.B., Picado-Muino, D. and Rosefield, M. (2008). "Information from inconsistent knowledge: A Probability Logic approach", Van-Nam Huynh et al (eds.) *Interval/Probabilistic Uncertainty and Non-Classical Logics, Advances in Soft Computing* 46: 291–307.

[Paris & Rafiee Rad 2008] Paris, J.B. and Rafiee Rad, S. (2008). "Inference Processes for Quantified Predicate Knowledge", W. Hodges, R. de Queiroz (eds.) *Proceedings of WoLLIC, LNCS* 5110: 249- 259.

[Paris & Rafiee Rad 2010] Paris, J.B. and Rafiee Rad, S. (2010). "A Note On The Least Informative Model Of A Theory", F. Ferreira, B. Lwe, E. Mayordomo and L. Mendes Gomes (ed.) *Programs, Proofs, Processes CiE 2010, LNCS* 6158: 342–351.

[Picado-Muino 2008] Picado-Muino, D. (2008). Doctoral Thesis, University of Manchester, UK.[see http://www.maths.manchester.ac.uk/ jeff]

[Popper & Miller 1983] Popper, K. and D. Miller (1983). A Proof of the Impossibility of Inductive Probability, *Nature* 302: 687–688.

[Priest 1979] Priest, G. (1979). "Logic of paradox", *Journal of Philosophical Logic* 8: 219241.

[Priest 1987] Priest, G. (1987), *In Contradiction*, Nijhoff.

[Priest 1989] Priest, G., Routley, R. and Norman, J. (1989). *Paraconsistent Logic*. Philosophia Verlag.

[Stalnaker 1968] Stalnaker, R. C. (1968). "A Theory of Conditionals", N. Rescher (ed.) *Studies in Logical Theory*, Oxford: Blackwell, pp. 98–112.

[Thagard 1993] Thagard, P. (1993). *Computational Philosophy of Science*. MIT Press/Bradford Books.

[Tversky & Kahneman 1974] Tversky, A. and D. Kahneman (1974). Judgment under Uncertainty: Heuristics and Biases. *Science* 185(4157): 1124–1131.

[Williamson 2010] Williamson, J. (2010). *In Defence of Objective Bayesianism.* Oxford: Oxford University Press.