The Bayesian Approach to Informal Argumentation: Evidence, Uncertainty and Argument Strength

Adam Corner

MSc Social Science Research Methods
University of Wales, Cardiff
BSc Psychology (Hons)
University of Wales, Cardiff

UMI Number: U584330

UMI

Dissertation Publishing

ProQuest

Acknowledgements

Firstly, I would like to thank my supervisor, Dr Ulrike Hahn, for her consistent and enduring guidance – I consider myself extremely lucky to have had such stimulating intellectual supervision, and empathetic personal support.

Secondly, thanks to my parents. They have encouraged and supported me in whatever I have done. They brought me up with an enthusiasm for learning, and a confidence in myself that has always stayed with me.

Finally, I would like to dedicate this Thesis to Danielle and Frankie. I have learnt more from you two than I could from studying for a thousand degrees. You are both always in my heart.

Thesis Summary

The work in this thesis contributes towards answering a simple, important and longstanding question: How do people evaluate informal arguments?

In Chapter 1, I review existing approaches to informal argumentation, and suggest that the Bayesian approach provides the most appropriate way of capturing informal argument strength. The Bayesian approach assumes that arguments are composed of claims and evidence. When people evaluate informal arguments, they make a probabilistic judgment about how convincing it is – that is, how likely the claim is to be true given the available evidence. The Bayesian approach is *normative*, because it makes predictions about how convincing different arguments *should* be. In Chapter 2, I examine the Bayesian claim to provide normative guidance for argument evaluation, and conclude that it provides solid normative principles on which to base an account of informal argument strength. The remainder of the thesis comprises experimental work in two distinct but related domains – the evaluation of socio-scientific arguments, and the evaluation of slippery slope arguments.

Understanding the public response to scientific messages about, for example, climate change, is becoming increasingly important. In Chapter 3, I report the results of four experiments (Experiments 1a, 1b, 1c, & 1d) designed to establish whether there are any differences in the way that people evaluate arguments about scientific topics as opposed to non-scientific topics. The data suggest that both scientific and non-scientific arguments are evaluated in a way that is broadly consistent with the rational predictions of Bayesian theory. In Chapters 4 and 5, I tackle a longstanding philosophical puzzle – when, if ever, is it rational to be persuaded by slippery slope arguments? Using Bayesian decision theory, and by identifying a mechanism on which evaluation of these arguments may be predicated, I demonstrate when and why slippery slope arguments are convincing (Experiments 2 – 9).

Finally, in Chapter 6, I conclude that the Bayesian approach provides a valuable metric for studying the evaluation of informal arguments, and identify some outstanding questions raised by my research.

Contents

Chapter 1 – Introduction

1.1   Chapter Overview

The work in this thesis contributes towards answering a simple, longstanding, and

important question: How do people evaluate informal arguments?

The purpose of this chapter is to set out how I have attempted to answer this question.

In doing so, I will foreshadow many of the issues that I later discuss in much greater

detail. Because the work contained in this thesis draws on a broad range of literatures,

and diverse empirical phenomena, my main goal will be to briefly describe each

chapter, outline the motivation for the work contained in them and summarise their

main findings. I will also introduce the theoretical framework I have used – the

Bayesian approach to argumentation – in some depth. First, however, I will provide

some context for the research reported in this thesis, outlining *why* the question of

how people evaluate informal arguments is a longstanding and important one.

1.2   Argumentation

Since antiquity, people have been fascinated by arguments – the tools of reason.

Many hundreds of years have elapsed since Aristotle produced his treatise on

argumentation, *De sophisticis elenchis* (Aristotle, 350BC), and later, his works on

syllogistic logic. Aristotle's desire to explicate the rules of rational engagement, and

to formalise standards of validity for deductive and inductive inferences, was a project

that many subsequently took it upon themselves to continue. The study of

argumentation has been central to philosophy ever since. One particular class of arguments that Aristotle introduced – *fallacies* – have particularly intrigued philosophers. Fallacies are typically defined as arguments that might seem convincing, but shouldn't be (see Hamblin, 1970), and are a long-standing puzzle in the philosophical literature on argumentation. Attempts to explain *why* classic argument fallacies such as the 'argument from ignorance', or circular arguments are fallacious have played a central role in developing a theory of argument strength in general – because in order to have a theory of why some arguments are weak, it is necessary to have an account of why others are strong (Hahn & Oaksford, 2006a, 2007a; Siegel & Biro, 1997; van Eemeren & Grootendorst, 2004; Walton, 1995). Indeed, an in-depth analysis of one so-called fallacy – slippery slope arguments – is a significant part of the empirical work presented in this thesis.

Historically, the only normative tools available for investigating argumentation – and therefore for distinguishing valid arguments from fallacies – were those of formal logic. In fact for a long time, formal logic and argumentation were inseparable (Hamblin, 1970), and arguments were measured by their formal logical validity. Basic logical rules such as *modus ponens* provide elementary guidance for assessing the validity of deductive statements, such that;

If it is raining (A), the pavement will be wet (B).

It is raining (A)

>>

Therefore the pavement is wet (B).

is a valid inference by *modus ponens*, whereas;

> If it is raining (A) the pavement will be wet (B).

> The pavement is wet (B)
>
> >>
>
> Therefore it is raining (A)

is the fallacy of *affirming the consequent*. Examples such as these are ubiquitous in introductory logic and critical thinking textbooks (Woods, Irvine & Walton, 2004).

Despite the historical dominance of logic as a means with which to evaluate patterns of formal argumentation (see, e.g., Gamut, 1991), there has been a great deal of criticism of formal logic as the appropriate normative system for assessing informal argument acceptability (e.g., Hamblin, 1970). This is partly because systems of logic can only assess arguments on the basis of their logical form, rather than their specific content, and therefore struggle to capture the richness of informal argumentation in natural language. Unlike the fallacies of formal logic, when informal arguments go wrong it is not clear that they violate logical norms in a straightforward sense (Hahn & Oaksford, 2007a; Ricco, 2007; van Eemeren & Grootendorst, 2004; Walton & Woods, 1989). For example, in their discussion of circular arguments – typically taken to be a fallacious form of argument – Hahn, Oaksford and Corner (2005) highlighted a fundamental problem with using logic to explain circular arguments' acceptability. Consider the following, frequently cited, example of the 'fallacy' of circular reasoning:

(1) God exists because the Bible says so and the Bible is the word of God.

Researchers have struggled to explain the fallacies generally, but circular arguments have been particularly troublesome, because this 'fallacy' typically embodies a deductively valid inference:

(2) God exists, because God exists.

Intuitively, this argument is unacceptable. Logically, it is an example of perfect deduction, as not only does the premise entail the conclusion – the premise *is* the conclusion. The problem is neatly summarised by Govier (1987):

"Many arguments which beg the question are formally valid, and in some what explains their begging the question is the very same thing that makes them formally valid: they contain a premise which is logically equivalent to their conclusion" (Govier, 1987, p177).

There exist, therefore, arguments that are logically valid but that are regarded as informally unacceptable. Because of this discrepancy between logical validity and informal acceptability (and a wide variety of other considerations), there has been a widespread philosophical rejection of formal logic as providing either necessary or sufficient criteria for evaluating informal argumentation (Boger, 2005; Hamblin, 1970; Heysse, 1997; Johnson, 2000). This has been buttressed by mounting psychological evidence that people do not naturally or consistently reason according to the rules of formal logic (e.g., Evans, 2002; Oaksford & Chater, 2001).

4

The rejection of formal logic as an appropriate system with which to judge informal arguments has led to two broad approaches to studying argumentation. The first, known as *pragma-dialectics* (van Eemeren & Grootendorst, 2004; but see also Walton, 1995, 1998), holds that the problem with logic as a normative standard for argumentation is its inability to account for the many factors that influence the acceptability of arguments in a dialectical context.

1.2.1 *The Pragma-dialectical approach to argumentation*

Van Eemeren and Grootendorst (2004) provide a useful definition of argumentation, which suggests why they consider logic to be insufficient for evaluating informal argumentation:

"Argumentation is a verbal, social, and rational activity aimed at convincing a reasonable critic of the acceptability of a standpoint by putting forward a constellation of propositions justifying or refuting the proposition expressed in the standpoint" (van Eemeren & Grootendorst, 2004, p1).

This definition of argumentation is *dialectical*, because it assumes that argumentation takes place between two or more people. It is *pragmatic*, because it seeks to elucidate normative standards for evaluating argumentation in the social norms and unwritten rules that govern argumentative discourse. According to van Eemeren and Grootendorst, arguments are tactical moves in a discussion, and are bound by rules that determine acceptable argumentative strategies. Similarly to linguistic theories of conversational competence, such as Grice (1975), the pragma-dialectical approach

holds that an argument is acceptable if it abides by the rules of a 'critical discussion'. For example, van Eemeren and Grootendorst claim that participants in a critical discussion should not be "confusingly ambiguous" (Rule 10) or present an attack that does not "relate to the standpoint that has been advanced by the other party" (Rule 3). On this account, acceptable arguments can be distinguished from fallacies by reference to a list of procedural rules.

By invoking practical rules such as this, pragma-dialectical theory takes the dialectical structure of argumentation as the point of departure, and develops a theory of argument acceptability from there. This can be directly contrasted with formal logic, which defines argument validity as a feature of the *argument itself*, and is insensitive to changes in linguistic or social context. In fact, one of the enduring achievements of the pragma-dialectical movement has been to shift the focus of argumentation away from the logical form of arguments, and on to their use and acceptability in practical situations. Particular forms of argument may therefore be acceptable in one context and unacceptable in another. As Walton (1995, 1998) has repeatedly stated, if arguments are fallacious, it is because they are a failure of communication, not a faulty form of inference.

This context dependency is illustrated by returning to the so-called 'fallacy' of circular reasoning. While the argument (1) about the existence of God would not do much to sway a sceptic, some circular arguments can be very compelling. Consider the following example:

(3) This fossil is from the Neolithic age, because it was found in a layer of rock where many fossils believed to be from the Neolithic age have previously been found.

Geologists frequently judge the age of fossils by the type of rock formation they are discovered in, yet one of the primary clues to the age of a particular rock formation is the type of fossils that are discovered in it. This process of reciprocal 'bootstrapping' seems intuitively acceptable, and using the pragma-dialectical approach, Walton (2005, 2006) has consistently distinguished between 'vicious' and 'virtuous' circular reasoning, appealing to the context of the argument and the type of dialogue it occurs in as differentiating factors. According to Walton, a circular argument is only fallacious if it fails to convince an opponent in a dialogue of the truth of the contentious statement (e.g. that God exists). Using the assertion that God exists as evidence to support the conclusion that God exists has failed in its purpose of convincing the sceptic that God actually does exist. The purpose of a discussion about a geological find, however, is somewhat different. Participants in a dialogue about the age of a particular fossil are simply using all the available evidence to come to a reasonable conclusion – the fact that the nature of the fossil and the nature of the rock imply a mutual conclusion does not amount to vicious circularity on the pragma-dialectical account.

Despite their intuitive appeal and popularity, however, pragma-dialectical theories of fallacy have difficulty in explaining why, in addition to *contextual* variation in argument strength (i.e. why circular arguments can be viciously and virtuously circular), there is substantial variation in the perceived strength of arguments that differ only in their *content* (i.e. why some virtuously circular arguments are more

convincing than others). Consider these final examples (discussed in Hahn &

Oaksford, 2007a) of circular arguments, this time in a minimal dialectical context:

(4) A: God exists

   B: How do you know that?

   A: Because the bible says so, and the bible is the word of God.

(5) A: Electrons exist

   B: How do you know that?

   A: Because we can see 3cm tracks in a cloud chamber, and 3cm tracks in cloud

   chambers are signatures of electrons.

There are striking similarities in the dialectical form of the arguments, and the

reference to 'unobservable' entities. It is difficult to see how the dialectical context

has been altered yet even examples (4) and (5) seem to vary clearly in their strength.

An analysis in terms of procedural rules of discourse seems forced in this minimal

context of a simple argument that is not embedded in a wider dialogical exchange. In

pragma-dialectical theory, the same type of argument may be acceptable or

unacceptable in different contexts, but there is no capacity to evaluate the *strength* of

individual arguments. Ultimately, a theory of when and how arguments go wrong is

not the same thing as a theory of argument strength.

Moreover, establishing a violation of procedural obligations is itself dependent on an

assessment of argument strength; a particular claim can fail to meet a reasonable

burden of proof only *because* it is weak. Consequently, that procedural violation

cannot in turn be invoked to explain the argument's weakness (Hahn & Oaksford, 2007b). That pragma-dialectical theories should fail in this regard should not come as a surprise. Procedural rules are just that – procedural – they are not rules for the evaluation of content. In law (the source of inspiration for many features of dialectical theories), procedural rules have an important role in enabling accurate, rational outcomes in legal domains – for example by disallowing confessions or testimonies obtained under duress. But they do not suffice to determine the actual outcome of a trial in and of itself. Crucially, we still have to evaluate the *content* of the evidence, not just the conditions under which it was obtained or presented. There is no reason why we should not seek the same standard of evaluation from a normative theory of argument strength.

1.2.2 The Bayesian approach to argumentation

A second approach to developing an account of argumentation uses Bayesian probability theory as a calculus for the evaluation of argument strength (Hahn & Oaksford, 2006a, 2006b, 2007a, 2007b; Oaksford & Hahn, 2004). In contrast to logic, which deals only with structure, and pragma-dialectics, which analyses arguments based on their contextual acceptability, the probabilistic, Bayesian approach is concerned directly with argument *content* (Hahn & Oaksford, 2006a, 2007a)

The Bayesian approach to argumentation starts from a very basic premise – that when evaluating the strength of an informal argument, one does so *probabilistically*. Informal arguments are typically comprised of claims (or hypotheses), backed by evidence. On the Bayesian account, people are assumed to ask a simple question

when evaluating arguments: How likely is a particular claim to be true, in light of the available evidence?

The extent to which an individual believes a claim to be true – that is, their *degree of belief* in a claim – is something that can be described probabilistically. A probability is simply a number between 0 and 1, where 0 indicates absolute certainty that the claim is false, and 1 indicates absolutely certainty that the claim is true. These probabilities need not correspond to any objective feature of the world, however – believing that there is a 0.2 probability of observing a black sheep amongst a white flock does not require you to have calculated or computed any kind of frequency or ratio of white to black sheep in the world. Your degree of belief is simply your *subjective* estimate that a particular claim is true. This basic and straightforward assumption about people's degrees of belief forms the basis for the Bayesian approach to understanding how people reason.

From this basic assumption that people's degrees of belief are fundamentally probabilistic, an enormous amount of ideas about formal and informal reasoning, hypothesis evaluation, evidential assessment and belief updating have developed. Howson & Urbach (1996), for example, have used the Bayesian approach to analyse the way in which scientists construct, test and eliminate hypotheses, design experiments and statistically analyse data. Legal scholars have shown how a probabilistic approach to evidence and uncertainty can be a valuable way of dissecting the complexities of courtroom testimonies, and of making an assessment of the reliability and credibility of witnesses (Schum, 1994). Epistemologists have used Bayesian principles to explain how people assess the coherence of sets of information,

and come to conclusions based on contradictory or disparate evidence. Most recently, there have been several attempts at applying the framework of Bayesian probability theory to the understanding of informal arguments – and, more specifically, at developing a rational theory of informal argument strength (Hahn & Oaksford, 2006a, 2007a; Oaksford & Hahn, 2004).

The Bayesian approach is named after Thomas Bayes, who developed a formula now referred to as *Bayes' Theorem*, which sets out the method by which an individual's degree of belief in a particular claim should be updated when they come across new evidence. Bayes' Theorem is a normative theory of belief revision, because it prescribes how our beliefs *should* change in light of new evidence. Essentially, the Bayesian approach holds that if we want our degrees of belief to be rational (i.e. in accordance with the laws that govern probabilities), then when we encounter new evidence, we should modify our beliefs in accordance with Bayes' Theorem. Whether we should, in fact, want our degrees of belief to be calibrated with the laws of probability is discussed in Chapter 2, but that the probability calculus provides a sensible metric for conceptualising belief revision is fairly uncontroversial[1]. For now, I will proceed on the assumption that the Bayesian approach provides a rational method for updating degrees of belief, in the light of new evidence.

The process of Bayesian belief updating is simple. For any claim (or hypothesis) *h*, one has a degree of belief (or subjective probability) $P(h)$ associated with it. Without looking out of your window, for example, you might have a degree of belief of 0.5

---

[1] The Bayesian approach is the dominant, but not the *only* mathematical theory of uncertain reasoning. Dempster-Shafer theory, for example (Howson & Urbach, 1996), focuses on 'belief functions' rather than simply degrees of belief, whereby a proportion of belief may be withheld altogether – that is, not assigned to any hypothesis (see also Parsons, 2001, for other measures of uncertain reasoning).

that it is raining, that is $P(h) = 0.5$. In Bayesian terms this is known as your *prior degree of belief* – as it is the degree of belief you hold *prior* to receiving any evidence. Imagine that you now encounter some new evidence e – droplets of water cover your window, and you want to update your prior belief that it is raining. Bayes' Theorem provides a rational method of getting from your prior belief $P(h)$, to a *posterior* belief given some new evidence. This posterior degree of belief in the claim is a conditional probability – your degree of belief in the claim *given* the new evidence – written as $P(h | e)$.

To calculate your posterior belief $P(h | e)$ you must make some attempt to quantify how much impact the new evidence you have obtained should have on your existing belief. According to Bayes' Theorem, you need to estimate how likely it is that the evidence you observed (i.e. droplets on your window) would have occurred if your initial hypothesis that it was raining was true, as opposed to if it was *false*. Formally, these are written as $P(e | h)$ and $P(e | \neg h)$ respectively, and known jointly as the *likelihood ratio*. They are often described in terms familiar from signal detection theory (Green & Swets, 1966) – as the *hit rate* and *false positive rate*. The ratio of hit rate to false positive rate jointly determines the diagnosticity of the evidence – that is the conditional probability of the evidence depending on the truth of the hypothesis. In this example, it is clear that observing droplets of water on your window is far more likely given that it was raining $P(e | h)$. There could of course be some other explanation for the droplets of water – perhaps the sprinklers have been turned on, or the person who lives in the flat above you has been watering their plants. But the hit rate in this example seems higher (we are more likely to see droplets of water given that it is raining), and the likelihood ratio therefore favours $P(e | h)$.

In line with the intuition that observing droplets of water on your window should increase your degree of belief that it is raining, Bayes' Theorem provides an update rule that increases your posterior degree of belief that it is raining given that you have just observed droplets of water on your window. Bayes' Theorem is stated formally below in Equation 1:

$$P(h \mid e) = \frac{P(h)P(e \mid h)}{P(h)P(e \mid h) + P(\neg h)P(e \mid \neg h)}$$

(Eq. 1)

To paraphrase Bayes' Theorem in words, the formula states that your posterior degree of belief is a function of your prior belief, and your estimate of the strength of some evidence, normalised by the *total* probability of the evidence (i.e. the chance that you will observe droplets of water whether or not the hypothesis was true). The Bayesian approach provides a general tool for measuring *argument strength* – the better the evidence, the stronger the argument, and the more your belief should change.

In the context of research on attitudinal change, the interpretation of reasoning patterns as probabilistic changes in subjective degrees of belief has precedent in the works of social psychologists in the 1960s (McGuire, 1960) and 1970s (Wyer & Goldberg, 1970). These early attempts to import probabilistic and statistical theories into accounts of belief change and attitude formation were based on the assumption that the relations among beliefs obey the laws of probability theory, such that as one belief is altered, other beliefs must be modified to accommodate this change. These so-called 'probabilogical' models (McGuire, 1981) sought to predict how a person's

belief in a particular conclusion would change when their belief in a related premise was revised. Specifically, people would be given a persuasive message and its impact on the belief it targeted was measured. The experimenter would then seek to ascertain the extent to which a further belief (one logically related to the first) had also changed – despite not being explicitly targeted by the persuasive message. The Bayesian approach shares the focus on belief and attitude change, but is focused, in the first instance, on the development of a theory that quantifies the direct impact a persuasive message *should* have in the first place, in contrast to the propagation of whatever change it happened to bring about throughout the wider attitudinal system.

Of course, arguments are not only about probabilities – many arguments pertain to the occurrence of a particular outcome. How desirable this outcome is – the utility associated with the outcome – is likely to have an important impact on how people evaluate the argument's strength. Consequentialist arguments, for example, take the generic form of 'if A, then B". The strength of a consequentialist argument will depend not only on how likely B is to occur, but also on how desirable B is. Stating that "If you mow the lawn, I will give you £100", is likely to be a more effective argument than stating "If you mow the lawn, I will give you £10", assuming that the probability of the two outcomes occurring is equivalent. How can considerations about outcome utilities be incorporated into a theory of argument strength?

Classical economic decision theory (von Neumann & Morgenstern, 1944) is a normative framework for decision-making in situations where outcomes are uncertain, based on the probabilities and utilities involved. According to this theory, agents should seek to maximize expected utility (i.e. potential gain) in their choices.

Decision theory has been enormously influential in both cognitive and social psychology. Bayesian decision theory (Edwards, 1961; Keeney & Raiffa, 1976; Savage, 1954) is simply the subjective interpretation of expected utility, such that instead of maximising utility according to some objective notion of value and probability, the probabilities and utilities are subjective. In the same way that Bayes' Theorem is a normative rule for updating subjective probabilistic beliefs, Bayesian decision theory is a normative rule for evaluating the potential courses of action available.

Both are termed 'Bayesian' because they involve a *subjective* evaluation of the facts. In the same way that believing that there is a 0.2 probability of observing a black sheep amongst a white flock does not require you to have calculated any kind of ratio of white to black sheep in the world, valuing one outcome over another does not mean that one outcome is better in any *objective* sense. This means that different agents can rationally choose different courses of action if their respective assessments of probabilities and utilities differ. Applied to the context of argument evaluation, this means that arguments can differ in strength depending on the audience to which they are addressed – a general characteristic of argumentation known as audience relativity (Perelman & Olbrects-Tyteca, 1969; van Eemeren & Grootendorst, 2004). However, there is still a normative standard in operation, in that the evaluation of decisions by a given rational agent must be derivable from more fundamental valuations, namely the probabilities and utilities they assign.

Specifically, the subjective expected utility of a decision (SEU) must correspond to the probability-weighted sum of the utilities associated with a particular course of

action. Equation 2 shows the SEU for any number of outcomes (x_i), where (P) is the subjective probability and (U) the subjective utility of each outcome:

$$\sum_i P(x_i)U(x_i) \quad \text{(Eq. 2)}$$

While Bayes' Theorem and Bayesian decision theory are related in the sense that they both involve applying a normative framework to subjective beliefs, they are mathematically and conceptually independent (although one could consider probabilistic beliefs about outcomes as a 'posterior' estimate of how likely an outcome is to occur, given the available evidence). In this thesis I will draw on both formalisations independently to motivate predictions about the strength of different types of arguments.

1.2.2 Argumentation: Summary

I have outlined three normative approaches to studying argumentation; formal (classical) logic, pragma-dialectical theory and the Bayesian approach. I have noted what are considered to be the shortcomings of the first two approaches in terms of providing a framework for studying informal argumentation, and introduced the Bayesian approach as a promising candidate for a theory of informal argument strength (Hahn & Oaksford, 2007a). I will now briefly describe how the rest of the thesis will proceed.

1.3   The Current Research

Throughout the previous section, I have referred to formal logic, pragma-dialectical theory and the Bayesian approach as being *normative* theories of argumentation. Establishing exactly what it means for a theory to be normative is, however, a non-trivial problem, and one that has puzzled philosophers for centuries (see, e.g., Railton, 2000). Matching actual human behaviour against putative norms has been at the core of research on judgment, decision-making, logical reasoning, and argumentation. But what are norms, and what makes them normative in the first place? In Chapter 2, I justify *why* the Bayesian approach provides a good normative theory of argument strength. Drawing on epistemology, and a typology developed in legal philosophy, I examine the question of how norms for behaviour in general might be justified, consider the normativity debate in argumentation, and ask how Bayesian norms for argumentation stand up as principles of argument evaluation. By comparing the normative basis for Bayesian, and pragma-dialectical theory, I propose that the Bayesian approach provides a solid normative basis with which to rationally evaluate informal arguments.

Having put forward my arguments for using the Bayesian framework as a normative theoretical approach, I then report a series of experiments designed to establish whether people are, in fact, Bayesian in their evaluation of informal arguments. I take a two-pronged approach, first conducting an investigation into a contemporary problem in argumentation, and then turning to a long-standing philosophical puzzle for the remainder of the experimental work.

In Chapter 3, I report the results of four experiments (1a – 1d) that focus on people's evaluations of arguments and evidence relating to contemporary scientific topics. Public debates about socio-scientific issues such as climate change are becoming increasingly prevalent. Scientific messages must be communicated to the general population, and public reaction to these messages in turn informs policy decisions. The public response to scientific messages about, for example, climate change, does not always seem to match the seriousness of the problem the scientists claim to have identified. Understanding how people interpret and evaluate science is therefore an important and pressing goal.

The literature on how people evaluate scientific arguments is disparate, and difficult to systematically integrate. And, in the absence of a coherent theoretical framework, it is difficult to pose questions about the relative strength of scientific arguments compared to non-scientific arguments. On the Bayesian account of argument strength, however, scientific arguments are simply arguments that *happen* to be about science. The key components that intuitively might seem to determine the strength of scientific arguments (e.g. how much evidence they contain, the relation of the evidence to the hypothesis, the reliability of the source reporting the evidence) have a simple interpretation in Bayesian terms – and are equally a feature of arguments about non-scientific topics. By bringing the Bayesian approach to bear on a number of different types of scientific arguments, I start to develop a framework for examining scientific arguments that allows systematic questions to be posed about the factors that influence their strength. Because the Bayesian approach is *normative*, it allows predictions to be made about when particular arguments should be strong and weak.

And because it is general and *content based*, it allows comparisons to be made between scientific and non-scientific arguments.

The results of these exploratory experiments suggest that although there are some interesting differences in how scientific and non-scientific arguments are evaluated, there is no reason to suggest that scientific arguments are 'special'. Rather, the variation in their perceived strength is attributable to the same factors that determine the strength of non-scientific arguments. Crucially, however, it is only the application of a normative framework for argument evaluation that permits questions about the relative strength of scientific and non-scientific arguments to even be posed in the first place.

Having introduced the Bayesian approach as a framework for studying a contemporary problem in argumentation, I then apply the Bayesian framework to a longstanding philosophical puzzle in Chapter 4: The slippery slope argument (SSA). The SSA has a bad philosophical reputation, but seems to be widely used and frequently accepted in many legal, political, or ethical contexts. SSAs warn against taking an initial action (e.g. legalising cannabis), on the grounds that it will cause the acceptability of some undesirable outcome (e.g. legalising cocaine) to be re-evaluated as positive in the future. In experiments 2-6, I show that SSAs that should be strong according to Bayesian criteria are perceived as more convincing, and bring about a greater degree of attitude change than SSAs that are weak in Bayesian terms. Specifically, I apply the framework of Bayesian decision theory, and analyse SSAs as arguments that vary in the probability and negative utility of their predicted outcome. The results indicate that the more probable and undesirable the outcome of an SSA is,

the more compelling people find the arguments to be. Chapter 4 provides an answer to the longstanding philosophical question of whether SSAs can ever be rationally strong arguments, as well as demonstrating that people's judgements of SSA strength are broadly in line with Bayesian predictions.

In Chapter 5, I extend my analysis of SSAs by identifying a well-known psychological mechanism on which people's evaluations of SSAs may be predicated – similarity based categorisation. In three experiments (7-9) I show that the more similar the 'initial action' and 'predicted outcome' of an SSA are, the stronger the argument is likely to be. If the beginning and end of a slope are alike, they are more likely to be perceived as belonging to the same category, making the slope seem *more slippery*. Thus, compelling SSAs may be based on genuine 'slippage' due to the inherent vagueness of many real world categorical boundaries – category membership is a dynamic process, and so presently 'unacceptable' outcomes may in the future be assimilated into an 'acceptable' category. In Experiments 7 and Experiment 8, I demonstrate the correspondence between evaluations of argument strength and categorisations judgements using a numerical measure of similarity – that is, a measure of similarity that is *objectively measurable*. That subjectively rational judgements of SSA strength correspond to objective measures of similarity suggests that SSAs are not simply fallacies that are 'wrong but persuasive'. In Experiment 9, I demonstrate a similar coupling of argument strength and categorisation decisions using materials that differ along a qualitative dimension of similarity.

In Chapter 6, I summarise the contribution that the research presented in this thesis makes to the literature on argumentation, identify some outstanding questions and areas for improvement, and suggest some possible directions for future research.

1.4   Chapter Summary

The central theme in this thesis is that the Bayesian approach permits vital questions to be formulated about both contemporary issues in argumentation (i.e. the interpretation and evaluation of socio-scientific arguments) and longstanding philosophical problems (i.e. when and how SSAs are convincing). In presenting the research I have conducted, I will draw on a genuinely diverse range of inter-disciplinary knowledge – including epistemology and legal philosophy, cognitive and social psychological accounts of reasoning and persuasion, decision theory and similarity, and the myriad of issues surrounding science communication.

Indeed, one of the novel aspects of the present research is that it utilises insights from a wide range of disciplines. Rather than extending an existing programme of research *per se*, I have endeavoured to study two phenomena (science communication and SSAs) that have an intuitive appeal, but have not necessarily received systematic empirical attention. Of course, while studying novel empirical phenomena has many benefits, it is a task that can only be sensibly accomplished using a sufficiently general theoretical framework. The message I hope to convey in the remainder of this thesis is that the Bayesian approach to argumentation seems to provide this.

Chapter 2 - Normativity and Argumentation: Why a Bayesian theory of argument strength?

"...to give up (normativity) appears to invite intellectual suicide (how could one recommend doing precisely that, for example?)" (Knowles, 2003, p33.)

2.1   Chapter Overview

In Chapter 1, I outlined in some detail the Bayesian approach to informal argumentation, describing it as a *normative* approach. Norms, that is, specifications of what we *ought* to do, have been central to the study of cognition whenever people have asked questions about human rationality. Matching actual human behaviour against putative norms has been at the core of research on judgment, decision-making, reasoning, and argumentation. But positing norms for studying argument evaluation pre-supposes that it is possible and desirable to establish norms for argumentation in the first place. In this chapter, before reporting any empirical data, I examine the question of how norms for argumentation might be justified. Drawing on epistemology, and a typology developed in legal philosophy, I consider the normativity debate in argumentation, and ask how pragma-dialectical and Bayesian norms for argumentation stand up as principles of argument evaluation. I conclude by claiming that the Bayesian approach provides a solid normative basis with which to evaluate informal argumentation.

2.2   Introduction

One of the central questions running through the experiments reported in this thesis is whether people are Bayesian in their evaluations of arguments – that is, do people approximate Bayesian norms in their evaluations of argument strength? The empirical data in this thesis therefore add to the enormous number of studies on judgment, decision-making, reasoning, belief revision and argumentation that compare *actual* behaviour to putatively rational *norms*.

Cognitive programmes (such as that initiated by Tversky and Kahneman investigating both the calibration and underlying mechanisms of people's 'intuitive statistics') place a huge emphasis on normative questions about behaviour. The seemingly considerable shortfalls in rationality exhibited by people's judgments that are apparent in phenomena such as the conjunction fallacy (Tversky & Kahneman, 1982, 1983) are topics of continuing interest to this day. Research has focussed on elucidating particular 'heuristics' and 'biases', but has also posed broad questions about the extent to which people are Bayesian – that is, the extent to which their behaviour is normative. In the equally sprawling literature on decision-making and rational choice, the norms of decision theory (in both its objective and subjective form) have guided the evaluation of decision-making behaviour in economic and psychological experiments (Edwards & Tversky, 1967; Pratt, Raiffa & Schlaifer, 1995). Much of the literature on human reasoning has focussed on logic, and human deviations from it. Studies of the Wason selection task (Wason, 1968), syllogistic reasoning (e.g. Johnson-Laird & Bara, 1984), and reasoning with conditionals (Evans & Over, 2004; Evans, Over & Handley, 2005; Johsnon-Laird & Byrne, 2002; Manktelow & Over,

1991; Oaksford & Chater, 2003) are ubiquitous in cognitive psychology – and usually motivated by a desire to document people's ability to reason according to putative normative standards.

Similarly, social psychologists of the 1960s (McGuire, 1960) and 1970s (Wyer & Goldberg, 1970) used logical and probabilistic norms to evaluate the consistency of beliefs, or measure belief change (Edwards, 1961; Slovic & Lichenstein, 1971). Despite pursuing an ostensibly descriptive agenda of outlining message effectiveness, contemporary theories of persuasion (Eagly & Chaiken, 1993; Kruglanski, Fishbach, Erb, Pierro & Mannetti, 2004; Petty & Caccioppo, 1984) regularly make use of 'strong' and 'weak' arguments to bring about attitude change. That some arguments are defined as stronger than others immediately raises the question of why this should be so – and suggests that normative questions about argument strength should occupy a central position in the study of persuasion.

Finally in the field of argumentation, both procedural (van Eemeren & Grootendorst, 2004) and epistemic (Biro & Siegel, 2006; Hahn & Oaksford, 2007a) normative theories have been put forward as the appropriate standard with which to assess people's use and acceptance of different types of argument. The role of normative considerations in studies of argumentation is explicitly acknowledged. Similarly to the many programmes designed to assess the development of 'critical thinking' in children (Kuhn & Udell, 2003), most argumentation theorists "…take it as obvious that the overarching goal of argumentation theory is the improvement of argumentation skills" (Gilbert, 2007).

In each of these areas, there is a recurring theme – are people rational?

In attempting to answer this fundamental question, the role of central norms has been twofold. First, the extent to which human behaviour matches up to these putative 'gold standards', and therefore the extent to which we might rightly claim to be rational, is of fundamental interest in its own right. It is also the question that has dominated the reception of this work in areas beyond psychology. Second, specific deviations have been critical in formulating and testing actual process theories of how humans go about these tasks (e.g., Kahneman and Tversky's Prospect Theory, 1979, or Johnson-Laird's mental models theory in the domain of logical reasoning, 1983).

At the same time, deviations from these supposedly rational standards have led to discussion about the standards themselves. In particular, spearheaded by Simon's notion of 'bounded rationality' (Simon, 1982), researchers have come to focus on the adaptive value of cognitive strategies as a normative standard (Gigerenzer, 1991; Gigerenzer & Selten, 2002; Gigerenzer & Todd, 1999).

On the one hand, this has led to the ever-increasing popularity of 'rational analysis' as a means for studying cognition (see Anderson, 1990, or Chater & Oaksford, 2008; Oaksford & Chater, 1998 for overviews). Here, an optimal computational solution to an environmental problem faced by an organism is identified, and provides a functional explanation of the organism's actual behaviour which is viewed as an approximation to that strategy. This framework has now been applied far beyond the reaches of judgment, reasoning, or decision-making, thus broadening the issue of rationality to novel domains such as memory (Anderson & Schooler, 1991) or

categorization (Anderson, 1990; Lamberts, 1995). On the other hand, the emphasis on adaptive value has also led some to question the normative status of probability, logic and decision theory as appropriate standards of rationality at all (Kahneman & Tversky, 1979; Noveck & Sperber, 2004; Stich, 1990).

However, given their central role, there has been very little discussion (in psychology at least) of *why* these norms should be considered normative – typically, putative norms are simply assumed to be normative (settled by work in other areas, such as philosophy), or denied outright (see, e.g. Bishop & Trout, 2005). At the same time, some have gone so far as to suggest that it is impossible to demonstrate human irrationality using the experimental method at all (Cohen, 1981), or that norms should be abandoned in the study of reasoning altogether (Elqayam, 2007).

It seems fair to say that the rationality debate is alive and well after nearly half a century of empirical evidence. Whether the right norms are being invoked in experiments, whether people adhere to these norms, whether people *should* be adhering to these norms, and what people's behaviour in reasoning experiments tells us about human rationality are all questions that are still up for grabs. But conducting research into people's ability to reason in line with given norms of rational inference presupposes something very important – that it is *possible* to derive norms for reasoning at all. What are norms for reasoning, and how might they be justified?

In this chapter, I take a step back from empirical debates about human rationality, and ask instead what it means for something to be considered normative. To this end, I first consider the question of how norms might be identified in general. I introduce

insights from discussions of normativity found in legal philosophy, and epistemology. Law as a body of norms, and theoretical insights from legal philosophy, turn out to be a very useful source here. I then consider the kinds of foundations for norms that legal philosophers and epistemologists have posited, as a means of granting normative status in the realm of reasoning. Finally, I examine evidence from the Bayesian and pragma-dialectical theories of argumentation in relation to their philosophical positions on normativity. I propose firstly that it is possible and desirable to invoke norms for rational argumentation, and secondly that a Bayesian approach provides solid normative principles with which to do so.

2.3   The Normative Question

What makes something normative? This question, or variants of it, can be found in extremely diverse fields of enquiry such as ethics, epistemology and legal philosophy, and capturing the notion of normativity is in itself a non-trivial task. The dictionary definition of "normative" is unhelpfully self-referential; "Of, relating to, or prescribing a norm or standard" (OED, 2006). Slightly more useful is the etymology of the word "norm", which derives directly from the Latin *norma*, the term used to describe a builder's square (Railton, 2000). The purpose of a builder's square is to allow *actual* cuts to be compared to an objective standard of correctness (i.e. a geometric right angle). When deviations are noted between the *actual* cut and the *norma*, corrections are made to the cut rather than the tool. The tool provides, therefore a normative standard with which to evaluate the cut, and epistemological, moral, or legal norms provide standards with which to evaluate human behaviour. However, the analogy really only captures some of what makes normativity such a

controversial philosophical topic, particularly when applied to human rationality. What if instead of a straight cut one wished to evaluate a curve? Now the tool seems distinctly inappropriate, and doesn't seem to provide normative guidance at all. In doubt is the applicability of the norm, which demonstrates that norms must somehow be derived – and that it is possible to derive the wrong ones.

In fact philosophers have recognised for a long time that simply observing whether a particular behaviour matches some prescriptive standard is not all that is required to understand the concept of normativity. If it were, then there would be apparently normative behaviour all around us. Driving on the correct side of the road in a foreign country is an example of something that we *ought* to do. If mammals do not lay eggs, and horses are mammals, then it *ought* also to be the case that horses do not lay eggs. But somehow, neither of these candidate conceptions of normativity seems to capture the essence of what it means for something to be normative.

On the one hand, the side of the road that people drive on in any particular country is fairly arbitrary – and would seem to confer compliance (for personal safety) rather than normativity. If normativity in rational argumentation could be equated with a shared set of agreed procedural rules, then "being rational is like a musician being in tune...all that matters is that we reason harmoniously with our fellows" (Chater & Oaksford, 2000). On the other hand, that horses do not lay eggs necessarily follows from the fact that mammals do not lay eggs, and this also seems to eliminate the need to label the statement normative. As Railton (2000) puts it, "If a normative *must* is to have a distinctive place in the world, then it cannot be the *must* of...conceptual necessity". Generally, then, we consider something normative if it adheres to some

(normative) standard of *what ought to happen*, rather than what *might* happen or what *must* happen.

2.3.1 Epistemology and Normativity

Assuming that we could settle for a definition of normativity that is neither post hoc and arbitrary nor simply subsumed by necessity, where might this notion of normativity come from – that is, how can norms be derived at all? Epistemologists have struggled for centuries with the question of what makes something normative, and it is still disputed to this day. Indeed, a not uncommon view is that *any* attempt to provide normative justification for beliefs is doomed to failure by one of three routes (known as the 'Munchhaussen Trilemma' – Siegel & Biro, 2008): (1) invoking an infinite epistemic regress (whereby A is justified by B, which is justified by C...etc); (2) developing a viciously circular series of arguments that depend on each other for their validity (e.g. normative beliefs are those that are justified – and justified beliefs are normative); or (3) introducing an arbitrary point at which a belief is simply declared 'justified', and the search for further justification aborted.

The notion of an infinite epistemic regress can be illustrated nicely by imagining a persistent child, who refuses to stop asking the question 'why?' While most parents will be familiar with such a sophisticated philosophical tactic, few will realise that the child is unintentionally tapping into one of the hardest questions in epistemology – at what point does something become *just so*? Some things do seem to possess this property of self-evidence – analytic truths, for example, which are simply true by definition. Epistemologists have been cautious, however, in granting the concept of

self-evidence legitimacy beyond a selective group of logical and mathematical

principles (Finnis, 1980) and it remains therefore a controversial philosophical

property. Outside of this general acceptance that some (very basic) knowledge is

simply self evident, there is much debate about how the rest of our beliefs can be

justified (Audi, 2002; Siegel & Biro, 2008), although it is possible to identify two

broad approaches.

The first is known as *foundationalism*. This holds that if we know anything at all, we

must have at some point derived 'direct' knowledge (often held to be sensory

information, although it is well known that sensory knowledge is 'constructed' as

much as it is 'perceived'). Foundationalists suggest that our knowledge and beliefs

must be anchored in something more concrete – some kind of directly observed

evidence. So, we can trace our knowledge about when to cross the road back to

sensory knowledge about speed, depth and colour cue perception.

The opposing view is *anti-foundationalism* (or *coherentism*) – whereby beliefs are

justified if they are coherent with the other beliefs that an individual holds. On this

view, holding a justified belief is "more like answering a question in the light of a

whole battery of relevant information than like deducing a theorem by successive

inferential steps from a set of luminous axioms" (Audi, 2002, p196).

Clearly, the issue of normativity and justification in epistemology is a fundamental

one, and epistemology offers a rich source of information from which to draw insights

about normativity in argumentation. However, the epistemological approach to

normativity tends to operate at a fairly high level of abstraction (typically concerning

beliefs and knowledge), whereas argumentation and reasoning theorists seek to assess *specific* norms, and how they might be justified as principles of rational debate. Because of this discrepancy in analytical approach, I have also used legal philosophy to guide my analysis of normativity. Legal philosophers have sought to deal not only with the wider question of whether universal norms are possible and how they might be discerned, but also the more immediate question of how an *individual* rule contained in a specific legal system acquires normative status. This fits with my goal of taking suggested norms for argumentation and considering how their normative status might be founded.

2.3.2 Legal Philosophy and Normativity

Several broad strategies for bestowing normativity can be found in legal theory. These strategies map closely on to the foundationalist/coherentist distinction that epistemologists have pursued.

The first strategy seeks to derive normative status from other norms (Kelsen, 1941), and is analogous to the foundationalist view in epistemology. On this account, normative power is derived from deferral to ever more fundamental and *a priori* truths. Needless to say, many epistemologists (e.g., Railton, 2000) consider this account to be highly problematic, as derivation of normativity from other norms faces the problem of a potentially infinite regress. Kelsen seeks to avoid the regress problem through the adoption of a single, otherwise content-less basic norm ("*Grundnorm*") that he claims must form the underlying basis for a legal system. His theory is an attempt to find a point of origin for *all* law, on which basic legal

principles (and the specific laws that derive from them) obtain their legitimacy. However, short of positing some sort of *a priori* 'super rule' that could infuse attempts at knowledge acquisition with a normative seal of approval, many epistemologists (e.g., Railton, 2000) have rejected the idea that normativity is a feature of the world in some immutable sense.

A second strategy, analogous to the epistemological approach of coherentism, seeks to derive normativity from assent or recognition (Hart, 1961). However, this strategy too has faced much philosophical criticism. Deriving normativity from recognition raises the question of when, if ever, the normative can be derived from the descriptive, that is, *ought* inferred from *is* (and vice versa, see e.g., Hume, 1740; for a discussion of the is-ought fallacy in the rationality debate, see Stanovich, 1999). Quite simply, from the fact that I *ought* to be at my desk, it does not follow that I actually *am* at my desk. From the fact that I *am* at the bar, it does not follow that I *ought* to be there. Breaching the ontological divide between *is* and *ought* means that norms are permitted to be normative simply because they are the norms that we follow. Blending the descriptive with the normative in this way would seem to drain normativity of its coercive power – if we can only say that one norm is as good as another so long as it is agreed upon, then it becomes difficult to evaluate behaviour as right or wrong in any meaningful sense.

Hart (1961) responds to this criticism in two ways. First, as a legal positivist, the normativity he is defining for the legal system is not an absolute one in the sense of immutable universals. Like anti-foundationalist epistemologists, Hart is not seeking to derive truths with a single point of origin, and law is viewed as separate from

morality. Assent or recognition seeks to define only a qualified obligation for those that fall under the scope of this assent. At the same time, however, theft does not simply become legal for those who want to steal because it is a qualified assent that counts; specifically it is the recognition of a particular group – the officials administering the system – that counts. This strategy of deferring to an 'expert' as a method of protecting normativity against accusations of arbitrariness has also been entertained in relation to norms for reasoning (Stich, 1985, 1990; Stanovich, 1999; Elqayam, 2003), and as we shall see, theories of argumentation.

Having identified the different notions of normativity that have developed in epistemology and legal philosophy, I will now focus on two competing normative theories of argumentation – pragma-dialectical theory and the Bayesian approach – and apply the broad strategies that epistemologists and legal philosophers have identified to their candidate norms of argumentation.

2.4 Normativity and Argumentation

As discussed above, standards of rational inference have been a topic of interest since antiquity. For much of this time, logic (in one form or another) has been the putative standard against which arguments are evaluated. However, there has been an increasing perception fuelled by a wide variety of considerations that logic cannot provide an appropriate standard by which to judge argument strength (see, e.g., Boger, 2005; Hamblin, 1970; Heysse, 1997; Johnson, 2000; see also Oaksford & Chater, 1991; 1998; Evans, 2002 for critiques of logicism in the field of reasoning more generally).

As discussed in Chapter 1, this has led on the one hand to a dialectical (or rhetorical) approach to understanding argumentation (Toulmin, 1958; van Eemeren & Grootendorst, 2004; also Slob, 2002 for discussion), based on the assumption that the problem with logic as a normative standard for argumentation is its inability to account for the myriad of practical influences on the acceptability of arguments in a dialectical context. On the other hand, it has led to the rise of Bayesian probability as an alternative calculus for the evaluation of argument strength (e.g., Hahn & Oaksford, 2007a). Here, the problem with classical logic is the imposition of a binary normative standard that permits argumentation to be only valid or invalid – and nothing in between. Moreover, logical inference is fundamentally about truth preservation, rather than capturing *changes* in beliefs (Hahn & Oaksford, 2007a). Even more fine grained and multi-valued logics do not therefore get at the heart of the problem of informal argumentation: How does new evidence impact on existing beliefs?

Hence, there are now two complementary sets of purportedly normative theories of argumentation: *Procedural* theories that propose normative rules for dialogical exchange, such as pragma-dialectical theory (e.g., van Eemeren & Grootendorst, 2004; but also Alexy, 1989) and attempts to establish a normative epistemic framework for the evaluation of argument *content*, such as Bayesian theory (Hahn & Oaksford, 2007a, but also Korb, 2004, and Goldman, 2003). These two types of normativity in argumentation map well onto the legal and epistemological typology identified above – pragma-dialectical theory attempts to derive norms from assent or recognition (i.e. coherentism), while the Bayesian approach seeks to ground standards of rational argument in axiomatic mathematical principles (i.e. foundationalism).

2.4.1 Pragma-dialectical Normativity

As outlined in Chapter 1, van Eemeren and Grootendorst (2004) claim that argumentation must be seen as a social act, designed to resolve a difference of opinion. They emphasise the idea of an idealised model of critical discussion, a method by which speech acts can be critically evaluated in an argumentative discourse. Van Eemeren and Grootendorst propose that a series of social norms and unwritten rules (closely related to linguistic theories of conversational competence, e.g. Grice, 1975) patrol the boundaries of argumentative acceptability. A sample rule is as follows:

Rule 2 states that

"the discussant who has called the standpoint of the other discussant into question in the confrontation stage is always entitled to challenge the discussant to defend this standpoint" (van Eemeren & Grootendorst, 2004, p 137).

These procedural rules of conversation are analogous to Hart's (1961) strategy of deriving legal norms from prominent existing legal conventions. From the consideration of the legal example, several questions immediately arise; whose assent is relevant here, what kind of normativity is granted, and to whom does it apply? Social conventions have developed historically, and it is not clear that our pragmatic rules for discourse are any more immutable than our conventions governing clothing or politeness. In other words, the normative status that pragma-dialectical rules possess cannot simply be assumed to be of a universal nature. As in the assent-based

approaches to legal rules, however, such a lack of universality does not mean that assent *cannot* bestow normativity on conventions as we currently find them (though if does raise awkward questions regarding the conditions under which these rules may change, analogous, again, to the legal situation).

Anti-foundationalist (or coherentist) epistemologists maintain that epistemological norms are not derived from *a priori* sources, but rather develop indefinitely, in the same way that the Kuhnian notion of paradigmatic science does (Kuhn, 1970). Similarly to Kuhnian science, however, the fact that the norms are subject to temporal change does not invalidate their normativity, or bestow an unacceptable degree of relativism. Anti-foundationalists argue that as the quest to understand the natural world will never be fully completed, even our 'best' epistemological norms will ultimately be replaced someday. But equally, "since there is not any question of transcending the situation we are in at any time, there is no perspective from which we can regard them as only relativistically valid" (Knowles, 2003, p67). For anti-foundationalist epistemologists then, the paradox of deriving norms from assent is less problematic in reality than it is in theory – epistemological norms may be mutable, but on this account they are not conceived of as arbitrary, or weak and relativistic.

It is unclear, however, whether a similar case can be made for procedural theories of argumentation. Here we find a similar conception of normativity, but it is not at all obvious that an individual (or society) is incapable of 'transcending' a conversational context. In fact, it is straightforward to propose alternative procedural rules that although unfamiliar, could plausibly have developed through a process of assent. For example, what if there was a rule requiring that in disputes involving more than two

people, an individual could only respond to the person who spoke immediately before them? Of course, it is possible to think of disadvantages to this rule, but it is equally possible to construct a case in its favour – the rule prevents confusion, promotes orderly conduct, guards against two points being discussed at the same time, etc. While it is almost impossible to transcend epistemological principles, the possibility of conceiving of worlds where the procedural rules of argumentation are radically different suggests that the stance of anti-foundationalists in relation to mutable norms is not really tenable for proponents of pragma-dialectical theory.

Perhaps more troubling for pragma-dialectical theory, though, is the *type* of assent that pragma-dialectical theory invokes. This is because there exists, in my view, an important distinction between *developmental* assent and *evaluative* assent. The pragma-dialectical conception of normativity expressed in the idea of the critical discussion is that normative models of argumentation are simply idealised expressions of individuated behaviour that have accumulated by *developmental* assent to become norms. This accumulation of norms is captured in the notion of an ideally rational 'reasonable critic' (who has internalised these accumulated norms, and can subsequently ensure that ideally rational rules of debate are respected), and by the development of lists of questions that this rational critic could use to distinguish 'acceptable' from 'unacceptable' arguments. The normative status of the pragma-dialectical approach is typically asserted, rather than derived, by its proponents (e.g., Hoeken, 2001a, 2001b; O'Keefe, 2005), and for pragma-dialectical theories of argumentation the situation is less straightforward than in law – in the legal example, one at least has some evidence that officials actually apply the rules in question. The rules identified by pragma-dialectical theory, are, at best, implicit in our day-to-day

discourse. It cannot therefore simply be assumed that the right rules have been identified, or, by consequence, that they are assented to in daily practice. What is really required is *evaluative assent*; that is, would most people agree that pragma-dialectical norms *should be followed*?

To date, there has been very little direct empirical assessment of people's understanding and agreement to these putative rules (for the exception see Bailenson & Rips, 1996; Rips, 1998; Rips, 2001), although there is some indirect evidence that people find arguments that observe simple dialectical principles such as clarity and explicitness to be more compelling (O'Keefe, 1997a, 1997b).

However, should these rules be viewed as prescriptive for all members of a community within which they dominate, or are they binding only to those who directly subscribe to them? In a framework such as Hart's, individuals cannot simply opt out, because it is only the recognition of a particular group backed by authority and sanction that bestows normativity. Some philosophers have suggested that recognition based normativity can be validated by deference to an 'expert source'. Stich (1985), for example, has suggested that rules of inference and epistemological principles may be justified by the process of *reflective equilibrium* carried out by a suitably expert source. Reflective equilibrium is simply the consideration (or evaluation) of an inductive process. What Stich's proposal amounts to is the suggestion that if an individual wished to establish whether a particular inductive principle was, in fact, normative, this could be derived from an assent based process so long as the assent comes from the reflective equilibrium of the individual's 'cognitive betters' – the experts in the particular inferential domain. However, Stich

himself notes that it is rarely obvious who the 'cognitive betters' in any given situation are, and it is not clear who the privileged group might be in the context of procedural theories of argumentation (although see Stanovich [1999], who suggests using measures of intellectual competence).

Van Eemeren and Grootendorst's (2004) reference to a 'reasonable critic', as a normative notion itself, somewhat begs the question of how this reasonable critic came to have normative authority – the locus of normativity is simply shifted elsewhere. In addition, empirical attempts to invoke pragma-dialectical criteria as a normative model of argumentation often require participants to be trained at length with complex evaluative criteria before they can perform the task (Hoeken, 2001a, 2001b). Bearing in mind that the notion of a 'reasonable critic' has been developed using an assent based process this seems a somewhat unreasonable expectation of what a reasonable critic should be capable of. The notion of normativity in the assent to social conventions therefore remains somewhat elusive.

2.4.2   Bayesian Normativty

By contrast, as an example of an epistemic theory of argumentation, stands the recent, Bayesian conception of argument strength. This account has been developed to provide a formal treatment of a range of classic *argument fallacies* such as the argument from ignorance, circular arguments or slippery slope arguments (Hahn & Oaksford, 2007a, but also Korb, 2004). Such a formal treatment has been a longstanding goal in fallacy research (Hamblin, 1970), and, by virtue of providing an

explanation of when particular arguments are weak, the account necessarily also provides an account of when arguments are strong.

As outlined in Chapter 1, on the Bayesian account of argument strength individual arguments are composed of a claim and evidence in support of that claim. Both claim and evidence have associated probabilities, which are viewed as expressions of subjective degrees of belief. Bayes' theorem then provides an update rule for the degree of belief associated with the claim in light of the evidence. Hence, argument strength is a function of the degree of prior conviction, the probability of evidence, and the relationship between the claim and the evidence – in particular how much more likely the evidence would be if the claim were true. In addition to theoretical analysis, there is also experimental work suggesting that people share the normative intuitions derived from the account (Hahn & Oaksford, 2007a; Corner, Hahn & Oaksford, 2006). This, of course, is the central theme pursued in this thesis.

Bayesian probability theory shares conceptual ground with foundationalist epistemological approaches and Kelsen's (1941) proposal that there are fundamental norms, which bestow normative power independently of whether they are followed or not. Could the normative power of Bayesian theory be rooted in self-evident, foundationalist principles?

In the context of epistemology, such a possibility has been voiced explicitly by Knowles (2003). Knowles claimed that the only cogent response to the problem of an infinite epistemic regress is to maintain that norms at some fundamental level must be "self evident, indubitable, self demonstrating or something of that ilk" (Knowles,

2003, p15). Elaborating on this position, Knowles suggested that certain logical or mathematical principles might be good candidates for norms that are self-evident.

Bayes' theorem follows directly from the axioms of probability theory – indeed, Bayes' rule is a *consequence* of the Kolmogorov probability axioms (Korb & Nicholson, 2004; Schum, 1994). These axiomatic mathematical statements are extremely minimal, and provide only the most elementary normative guidance. They define probabilities as non-negative numbers between 0 and 1, state that the probability of a certain event must be 1, and stipulate that the joint probability of any mutually exclusive events is equal to the sum of their individual probabilities. From these three axioms the definition of conditional probability can be derived, from which Bayes' Theorem directly follows (see Howson & Urbach, 1996). Some statisticians have gone so far as suggesting that probability theory is the 'inevitable' (i.e. the only sensible) method of describing uncertainty – which makes it an attractive method for deriving normative standards in theories of rational argumentation and belief revision (Lindley, 1982).

However, such an appeal to self-evidence might be perceived to be a cheat. From probability theory's status as a mathematical object, it does not follow that its *application* to day-to-day inference is normative also. And that this application is not self-evident can be read off from the variety of alternative calculi that have been proposed to this effect (e.g., Dempster-Shafer theory, see Howson & Urbach, 1996 for critical discussion), and from the debate surrounding the Bayesian interpretation of probabilities as subjective degrees of belief.

2.5   The Dutch Book Argument

Perhaps the most famous argument in the literature on Bayesian normativity is known as the Dutch Book Argument (DBA). The DBA has served as the central normative justification for Bayesian theory since Ramsey first proposed it (1931; see also de Finetti, 1974). It is based on linking degrees of belief to the betting preferences of a rational person – that is, a person with (hypothetical) betting preferences that conform to the probability calculus. A Dutch Book is a combination of bets which can be shown to entail a sure loss. Like Bayes' Theorem, a Dutch Book is simply a mathematical statement, and is philosophically uncontroversial. The Dutch Book *argument* connects degrees of belief to a (theoretical) willingness to bet by assuming that a person with degree of belief X in a proposition P would be willing to pay up to £X to bet on P. Being Bayesian – that is, being in possession of degrees of belief that conform to the probability calculus – provides immunity from Dutch books. People who have degrees of belief that do not satisfy the Kolmogorov axioms can be made to suffer a sure loss in a betting situation, as it is possible to construct a Dutch Book against them.

Consider, for example, an individual who had a degree of belief of 0.51 that it was raining, but also a degree of belief of 0.51 that it was not raining. Their total degree of belief would exceed 1, violating an axiom of the probability calculus (the probability of a certain event – that is, either rain or not-rain – should equal but not exceed 1). If this individual's betting preferences matched her beliefs, she should be willing to bet £0.51 to a bookie's £0.49 that it was raining, but also willing to bet £0.51 that it was not raining (to a bookie's £0.49). Her total bet would be £1.02, but the most she could

possibly win would be £1.00 – guaranteeing her a loss of £0.02. As accepting odds that lead to a sure loss in a betting situation would seem to be uncontroversially irrational, the DBA offers normative justification for being Bayesian that is directly based on the axiomatic principles of the probability calculus, and an extremely basic set of conditions for economic rationality (i.e. never entertain odds where every outcome entails a loss for yourself). When translated into argumentation, the DBA simply ensures that reasoners do not have conflicting or inconsistent degrees of belief in a hypothesis.

It is worth noting that the DBA does not depend on anybody actually winning or losing money – or anybody even betting at all. As several authors, including Christenson (1996) have shown, it is the principle of being vulnerable to a sure loss that is the essence of the argument: "The argument's force depends on seeing Dutch Book vulnerability not as a practical liability, but rather as an indication of an underlying inconsistency" (Christenson, 1996, p455).

The simplicity and elegance of the DBA as normative justification for Bayesian theory has attracted much support (e.g. Davidson & Pargetter, 1985). However, as Armendt (1993) notes, one way of judging the significance of an argument is by "the number and variety of (attempted) refutations it attracts", and by this measure, the DBA is very significant indeed. As will become clear in the following section, many critiques and defences of the DBA exist. In reviewing the literaure on the DBA as normative justification for Bayesian theory, criticisms of the argument seem to fall naturally into two broad categories – those that posit caveats to its universality, and those that question whether the DBA does enough to justify its position as the central

determinant of Bayesian normativity. In what follows, I have used this typology to structure my examination of DBA critiques.

2.5.1 Criticisms of the DBA – type I

The first category of criticisms of the DBA are those that question whether the simplicity of the argument – that the reason one's beliefs should conform to the probability calculus is because this provides immunity from betting losses – really captures the range of situations we might encounter where belief coherence is important. Waidacher (1997), for example, argues that the DBA does not provide a foundation for normative theories of rationality because it only applies to situations with a particular formal structure – specifically, where there is a linear relation between degree of belief and payoffs. Unless we accept the "far-reaching and highly implausible hypothesis" that all the situations we face in our life can be faithfully modelled by this hypothetical structure, then the DBA provides normative justification only for a limited range of situations, and consequently is not enough to provide the basis for a normative theory of rationality. Similarly, Davidson and Pargetter (1985) note that the DBA is based on the assumption that all parties have equal access to knowledge about the outcomes of bets – whereas in reality, inequalities in informational access may exist.

There is a sense in which arguments such as these are ultimately only capable of adding caveats to the DBA – they simply cannot evaluate the DBA as a source of normative authority, or the link it provides between betting preferences and degrees of belief. Undoubtedly, it is *possible* to conceive of situations where one might wish to

accept a Dutch Book, or where betting preferences and degrees of belief are not systematically related. Perhaps I am eager to impress a new acquaintance, and consider the financial losses I incur in irrationally accepting bets with a guaranteed loss to be a small price to pay for their jubilant mood. But are these cases typical, or merely the exception that proves the rule? Waidacher's (1997) argument amounts to an objection about assuming the value preferences of agents, which Sibler (1999) dismisses as "misidentifying relatively superficial problems in the application of utility theory as potentially devastating flaws in its foundation" (Sibler, 1999, p249).

It would seem therefore that arguments such as these can be dealt with by adding a simple *ceteris parabis* clause to the DBA – all other things being equal, it is rational for your degrees of belief to obey the axioms of the probability calculus.

Other authors have suggested that there are situations where betting preferences and degrees of belief may diverge, and that therefore the DBA does not do enough to justify its position as the source of normative authority for being Bayesian. Kennedy and Chihara (1979) suggest that playing intentionally poor hands in a game of Poker (i.e. knowingly allowing Dutch Books to be made against you) may be a rational strategy in the long run, as it may convince your opponent that you are a less sophisticated player than you are. Having established a false sense of security in your opponent, by losing a series of small bets, you may then stand a better chance of winning a big pot of money later on (the well known technique of 'hustling'). An economic analogue of this betting strategy is the practice (commonly employed by large retailers) of running certain product lines at a loss. Having enticed customers in by selling some products at an unprofitable rate, they are more likely to sell products

on which they are making a profitable return. Kennedy and Chihara claim therefore that there may be situations in which it is rational to accept Dutch Books.

However, although Kennedy and Chihara's argument seems compelling, in my view they fail to identify the crucial feature of these 'long run' strategies: The only reason that small losses (i.e. minor violations of the probability calculus) can be permitted in the short term, is that larger profits (i.e. better than 'fair' bets) are ultimately achieved. Large retailers can only 'loss lead' on certain product lines because it is good for their business overall. The hidden assumption in Kennedy and Chihara's argument is that loss making strategies are *only* rational because at some point the pendulum will swing the other way, and your poker opponent (or the consumer) will be persuaded to accept worse Dutch Books than the ones you suffered. Local losses must ultimately be counterbalanced by global gains, or else the acceptance of Dutch Books can no longer be claimed to be a rational strategy.

It is not enough therefore to demonstrate that people might sometimes prefer to maximize other utilities. But it is also insufficient to point to evidence that people pursue what appear to be non-normative strategies. This is because norm or value conflict does not negate normativity. This is readily apparent in law, where rules are not without exception. The killing of another individual is prohibited and sanctioned in British law, yet there are several 'full defences' against a charge of manslaughter, such as using reasonable self-defence against an attacker. Despite these exceptions the norm clearly remains. In doubting the normative status of the rules governing manslaughter under British law one would have to show not just that there are exceptions to the rule, but that the normative power of the rule was consistently

challenged. The same level of refutation is required for normative theories of rational argumentation, and it is not clear that such a refutation can be formulated against the DBA as normative justification for Bayesian rationality.

2.5.2   Criticisms of the DBA – type II

The second type of criticism of the DBA is more substantive – that coherence with the axioms of probability is not in fact a necessary (or sufficient – see Rowbottom, 2007) condition of rationality. For example, Hajek (2005) has claimed that proponents of the DBA have ignored the logical compliment of the argument – that when probabilistic coherence is violated, you are equally as likely to accept a 'Good Book' as a Dutch Book (with a 'Good Book' being a set of betting preferences that guarantee you a sure win). Given that no-one would argue that accepting a sure win is irrational, how can probabilistic coherence be a necessary condition of rationality? Clearly, adding the extra assumption that we are more likely to encounter 'Dutch Bookies' than 'Good Bookies' (i.e. that people are more likely to take advantage of probabilistic incoherence than reward it) is unacceptable, as:

"Susceptibility to a Dutch Book is a dispositional property of an agent, one that she has independently of what other people are out there, and what other people are like – in fact there need not be other people out there at all" (Hajek, 2005, p143).

Hajek provides an answer to his own puzzle, however, by proposing that the traditional DBA should be modified, such that instead of positing that it is rational to accept fair, and *only* fair, betting quotients, the DBA should state that it is rational to

47

accept fair or *better than fair* (i.e. favourable) betting quotients. Stated in this way, the

DBA ensures that only sure-loss violations of probabilistic coherence are irrational,

and the normative power of the argument is restored.

Several authors (see below) have proposed, however, that it may still be irrational to

insist on adhering to criteria of probabilistic coherence. Sibler (1999) has suggested

that probabilistic coherence is not consistent with 'instrumental' rationality, in the

sense that:

"(attempting to)...become coherent merely because of the logical possibility of

becoming the victim of a Dutch Book would involve such extensive exploration of the

logical relations among one's beliefs that it might well prove counterproductive and

instrumentally irrational itself" (Sibler, 1999, p255).

Sibler's argument echoes strongly the claims made by proponents of the 'bounded

rationality' approach to human reasoning, which I discuss in detail below. What

matters in the present context, though, is not whether people actually *are* Bayesian,

but whether they *think they ought* to be. In other words, what is critical here once

again is *evaluative* assent. This is measured not by the number of people observing

the norm, but by the number of people agreeing that the norm is, in fact, a good one.

Behavioural data cannot impact on a norm's integrity *per se* (a similar claim has been

made by Cohen, 1981, in defence of human rationality in general); rather actual

behaviour is at best, a weak indicator of evaluative assent. I am not aware, however,

of any studies measuring people's acceptance of probabilistic consistency as an ideal

(although see Slovic & Tversky, 1974, for evidence that educating people about the axioms of rational choice does not necessarily encourage them to use them).

Supporting the view that behavioural data is only a weak indicator of evaluative assent, Hookway (1993) has claimed that there is no necessary link between the normative status of a principle and our adherence to it. For example, a group may unanimously agree that being honest or open-minded is a positive trait (and a norm to be followed), but still fail to be honest or open-minded. This is because;

"Possession of epistemic virtue depends upon the possession of skills and habits whose possession is largely independent of the recognition that some state is, in fact, such a virtue" (Hookway, 1993, p76).

In fact it is rather hard to imagine that people would not, in general, accept that consistency and the immunity from Dutch Books that the probability calculus conveys are minimal standards of rationality that they would *like* to comply with. Closer inspection of the normative basis of Bayesian theory reveals a composite notion that mixes both derivation and assent. The self-evident axioms of probability provide a non-arbitrary foundation from which normative constraints on beliefs can be derived, and assent (to avoiding Dutch Books) underwrites the normativity of maintaining consistent degrees of belief. Again, this raises the question of who assents, and what evidence there is for this assent. The simplest answer is that people agree for themselves; rationality is a matter of individual choice, and people are free to pursue irrational strategies if they so choose. All the DBA proposes is that a rational person should exclude the possibility of negative consequences (i.e. sure losses) whose

unacceptability seems universally recognisable – a recommendation with appeal that it is difficult to dispute.

Returning to Armendt's (1993) proposal that the number and variety of refutations the DBA attracts is an indication of its importance, one final comment should be added in defence of the DBA, and its normative status. Not all philosophers have been persuaded that it is as compelling as it appears to be. Armendt himself has questioned whether the assumption that bets in a DBA are value independent always holds (essential if the DBA is to proceed from the axioms of probability). Some (such as Bacchus, Kyburg & Thalos, 1990) start out with the explicit goal of destroying the DBA, but manage only to prove that there may (although they do not specify them) be other ways of conceiving of rational behaviour – that perhaps the DBA is not the only path to epistemic integrity. But does this make the DBA unacceptable as the normative basis for Bayesian rationality? Armendt provides a succinct repost to those who would prematurely abandon the DBA:

"Demands that we assume nothing and prove strong conclusions, however the demands are disguised, are unreasonable…(A)n appropriate response is to demand from the critics something better. A Bayesian's admission that his theory can be improved, seen in these terms, is not thereby an admission that the current theory is nonsense. And the fact that nobody can (correctly) prove something from nothing does not make every theory equally good or bad" (Armendt, 1993, p20).

The DBA covers a lot of ground using an extremely minimal set of assumptions. And it seems to be, to the best of my knowledge, an appropriate normative justification for Bayesian rationality.

2.6   Bayesian Normativity and Computational Limitations

The argument raised by Sibler (1999) that maintaining probabilistic coherence in our degrees of belief implies an 'ideally rational agent' echoes the 'bounded rationality' approach to human reasoning (Simon, 1982). According to the bounded rationality hypothesis, in order to completely absorb (and therefore act on) the statistics of the environment, it would be necessary to possess computational powers far in excess of the human brain. This is an argument against maintaining Bayesian principles as a theory of rational argument that goes beyond debating whether or not people's behaviour is actually Bayesian – proponents of bounded rationality claim that it is simply not practical to expect people to be *capable* of observing Bayesian norms – at least, not without assistance (Gigerenzer & Edwards, 2003). Proponents of bounded rationality argue that instead we use a series of 'fast and frugal' heuristics to approximate normative solutions. Rationality under this interpretation is relative to the performance limitations of the individual and the demands of the immediate environment, and cannot be captured by a theory that proposes absolute norms. Without disputing the authority of Bayesian norms, a proponent of the bounded rationality approach might suggest that in positing norms for probabilistic coherence that are *in principle* unobtainable we have simply selected the wrong ones. Instead, we should take environmental limitations into account, and settle for norms that give

/

us a kind of contingent optimality – rationality defined not just by normative ideals but also by cognitive constraints.

There is certainly an appeal to these arguments, and in the sense that models of bounded rationality offer methods of obtaining rational outcomes that do not depend on computational powers beyond our reach, they paint a picture of rationality that resonates with our intuitive notion of what is 'reasonable' to expect from even the most rational individual. But is this sort of rationality, no matter how 'reasonable', actually normative at all? While it may well be more sensible to calibrate your rationality to standards that are within your grasp, this does not, in and of itself, make these standards normative. This is because normativity is about obtaining the *right* answer, not simply an answer that is as close to correct as can reasonably be expected given the circumstances. If normativity could be defined in this way – as an adaptive response to whatever circumstances you may find yourself in – it would confer an undesirable level of situational specificity. That adaptivity and normativity are not equivalent can be demonstrated by considering mechanisms of biological evolution which are certainly adaptive, but cannot be said to be 'correct' or normative in any meaningful sense of the word. While adaptive evolutionary mechanisms may well produce characteristics or behaviours that appear to be normative, normativity is not a precondition of adaptiveness.

A simple example helps to further highlight the differences between 'bounded' rationality and normative rationality. Imagine you have been set a particularly difficult multiplication problem to solve in your head – say 3784 X 457. For all but the most gifted of mathematicians this calculation is too difficult to solve accurately

given reasonable time constraints. A proponent of bounded rationality might suggest, therefore, that the normative course of action given computational limitations and environmental constraints would be to round the numbers down to something more easily calculable – and it would be difficult to argue with the reasonableness of this suggestion. Two issues immediately arise, however, that suggest that arguing that such a solution is *normative* might be misguided. Firstly, some people are better at computing long multiplication than others. There does not seem to be any *a priori* method of establishing who has normatively opted to round the numbers to 3800 and 450, and who has lazily multiplied 4000 by 500 and clearly got the wrong answer. We seem to have committed ourselves to a definition of normativity that gives us no normative guidance whatsoever. Secondly, getting close to the right answer given situational constraints does not prevent us from wanting to know what the right answer *actually* is. Adopting a definition of normativity that operates according to bounded rationality actually prevents us from making a normative judgment at all. While describing human behaviour as being rational whenever it does the best possible job given all the constraints on its operation makes intuitive sense, we must refrain from calling this behaviour normative.

So, in the same way that behavioural data demonstrating non-compliance with Bayesian norms does little to damage their normative status (only their descriptive validity), the observation that it is not always practical to expect Bayesian norms to be maintained does not undermine the claim that they are still normative standards. As Chater and Oaksford (2000) note, optimal models in economics, animal behaviour, or psychology rarely assume that agents are able to find perfectly optimal solutions to the problems that they face. It is widely accepted across a wide range of disciplines

that there is no contradiction in positing normative standards that are difficult to adhere to, even in principle. But one might wonder whether people would assent to norms that were impossible to achieve, and perhaps it is here that bounded rationality can claim to provide more appropriate normative guidance. Should we consider things that are *impossible* to be normative?

The idea that possibility is a necessary precondition of normativity has been a feature of Western legal systems, in particular in the area of contract law (see e.g., Lando & Beale, 2000) since antiquity. "Impossibilium nulla est obligatio" means "there is no obligation to do impossible things", and is a fundamental principle of Roman law. However, the question of impossibility arguably does not even arise in the present context. Behavioural data may support the claim that people cannot *always* be Bayesian, but it certainly does not support the claim that people cannot *ever* be Bayesian. Even the most ardent anti-Bayesian would not suggest that being Bayesian is impossible – although it may be demanding in many practical situations (and there is a vast literature detailing the heuristics and biases people may bring to bear on situations such as these – see, e.g., Gilovich, Griffin & Kahneman, 2002). In much the same way that the multiplication problem described above may be so time consuming that few would attempt to provide an exact answer, it is certainly not impossible. Of course, one can imagine ever more complex calculations that are beyond any intellect. But even these calculations have a *correct* answer – and that answer is no less correct for the lack of a person who can accurately compute it. Similarly, Bayesian norms are no less normative for the lack of an individual who is completely probabilistically coherent in their beliefs.

Indeed, proponents of bounded rationality are keen to point out that given the right cognitive tuition (Gigerenzer & Selten, 2002) or environmental tools, the difficulty of complex tasks can be greatly reduced. Presumably, the provision of probabilistic tuition has no bearing on whether a norm is normative or not. Few would want to argue that the lack of a calculator would destroy the normative authority of the mathematically correct answer in the multiplication problem described above, and by the same token, few would seek to question Bayesian norms simply because there are practical constraints on how achievable they are.

The development of Bayes' nets (see, e.g., Pearl, 1988) as a tool for implementing Bayesian computations also suggests that being Bayesian may not be as difficult as it first appears. Bayes' nets illustrate the power of 'conditional independence' – the idea that probabilistic information is often relatively unaffected by changes elsewhere in a probabilistic network. This suggests that although global probabilistic consistency is daunting in principle, local consistency is considerably more manageable. This is because in reality, so much of the probabilistic information we process is *conditionally* independent. For example, you are unlikely to need to update your degree of belief in the prevalence of an obscure species of lizard in the light of recent information you received about rising levels of acidity in the Atlantic Ocean, even though this evidence potentially bears on the global ecological system. This is because lizard prevalence is likely to be conditionally independent of sea acidity, given more directly related evidence – for example, the temperature of their habitat. Within a range of habitat temperatures, the level of sea acidity is effectively 'screened off', and no probabilistic changes to your beliefs are necessary.

To question the normative status of Bayesian inference it would be necessary to demonstrate not only that someone's behaviour was not Bayesian, not only that on occasion that person would not choose to reason in a Bayesian fashion, and not only that on occasion that person would not be capable of reasoning in a Bayesian fashion, but that that person would dispute the normative claim of probabilistic consistency most of the time. Given the mal-adaptive consequences of such a stance, it seems rather unlikely that many such individuals can be found.

2.7   Chapter Summary

It is clear that the normative question is a complex one, and in this chapter, I have tried to identify what seem to be the crucial aspects of normativity to consider when developing a normative theory of argument strength. By drawing on ideas about normativity from legal philosophy and epistemology, I have sought to shed light on what sort of normative theory might be appropriate for argumentation, and what features such a theory might need to incorporate. Based on the evidence reviewed in this chapter, it would seem that there are certain features that good normative theories possess – and that the Bayesian approach seems to embody these features.

By combining the self-evidence of the axioms of probability theory with the minimal economic rationale of the DBA, Bayesian inference seems to be based on solid normative principles that are not vulnerable to the problem of infinite regress. Assent is required to the extent that reasoners must agree that in general, adhering to the DBA and benefiting from the protection it provides against inconsistency is a good thing. Whilst people may often (for any number reasons) deviate from the norms of

Bayesian inference behaviourally, it seems unlikely that they would dispute the normative rationale of Bayesian principles.

In the next three chapters of this thesis, I report the results of a series of experiments that suggest that across a variety of argument evaluation tasks, people are, in fact, able to evaluate arguments broadly in accordance with Bayesian norms for argumentation.

Chapter 3   Evaluating Scientific Arguments: Evidence, Uncertainty and Argument Strength.

"...(A)rgument is manifest in the establishment of scientific knowledge. Science is the product of a community and new scientific conjectures do not become public knowledge until they have been checked, and generally accepted, by the various institutions of science...The rational processes of argument are the foundations of these institutionalized practices" (Newton, 1999, p555).

"What is rarely appreciated by the public...is that rather than being a weakness, dispute lies at the very heart of science...Understanding argument, as used in science, is therefore central to any education *about* science" (Driver, Newton & Osborne, 2000, p301).

3.1   Chapter Overview

As I explained in Chapter 1, I have focused my empirical research on two distinct areas of argument evaluation, using the Bayesian framework as a tool for posing novel empirical questions. In this chapter I examine the interpretation and evaluation of arguments about socio-scientific issues.

Public debates about socio-scientific issues are becoming increasingly prevalent, but the public response to scientific messages about, for example, climate change, does not always seem to match the seriousness of the problem the scientists claim to have identified. In this chapter, I ask whether there is something special about appeals

based on scientific evidence – do people evaluate scientific arguments differently to non-scientific arguments? The existing literature on scientific argument evaluation is disparate and it is difficult to provide an answer to this question. In an attempt to develop a systematic approach for studying people's evaluation of scientific arguments, I apply the Bayesian framework to some key aspects of scientific argument evaluation. The Bayesian approach permits questions to be systematically posed about how people evaluate scientific arguments. It also allows comparisons to be made between the evaluation of scientific and non-scientific arguments. Across four experiments (1a-1d), I demonstrate that participants evaluate both scientific and non-scientific arguments in a broadly similar way. The results are necessarily tentative, as the question of how people evaluate scientific arguments is a broad one. Despite this, the exploratory data in this chapter would seem to have some interesting implications for the successful communication of scientific arguments. And, most importantly, the application of the Bayesian framework permits questions about argument evaluation to be systematically posed at all.

3.2   Introduction

Public debates about socio-scientific issues such as climate change are becoming increasingly prevalent, and in many ways, science no longer belongs to scientists alone. Of course, it is still scientists who conduct scientific experiments, and scientists who publish the results of their research in academic journals. Likewise, it is still scientists who peer-review each other's publications and scientists who debate the theoretical principles and empirical conclusions contained within. But bodies of

scientific evidence and the debates that surround them are no longer restricted to these circles.

Today scientific debates in the public domain are commonplace, with scientific developments debated by politicians, journalists, and citizens groups. Many of the most important decisions we make (as individuals or as a society) are rooted in our understanding and evaluation of scientific evidence, arguments and claims. The communication of messages about socio-scientific issues such as climate change is becoming a matter of some urgency. The Intergovernmental Panel on Climate Change (IPCC), for example, has recently reported that:

"Most of the observed increase in globally-averaged temperatures since the mid-20th century is *very likely* due to the observed increase in anthropogenic green house gas concentrations...(these are) expected to have mostly adverse effects on natural and human systems" (IPCC Synthesis Report, 2007).

The public response to messages such as this, however, does not always seem to match the seriousness of the problem the scientists claim to have identified. Academic interest in the public understanding of science driven by concerns such as these is rapidly increasing. This is reflected in the range of researchers – philosophers, social scientists, communication scholars, policy experts and science educators – who study the many aspects of science communication (Collins & Evans, 2007; Gregory & Miller, 1998; Pollack, 2005). Such diverse interest in the topic suggests that the process of communicating science to the general public is not straightforward. Is there something special about the way that scientific messages are evaluated by the public?

Could the communication of science be *improved*, if only we had a better understanding of the process? Certainly, improvement in scientific literacy is the explicit goal of science educators (see, e.g., von Aufschaiter, Erduran, Osborne & Simon, 2008), and the implicit goal of much of the work that has been conducted into the public understanding of science.

Despite the sustained attention that the issues surrounding science communication have received, a clear understanding of how people interpret and evaluate basic scientific messages not been forthcoming. In particular rather little is known about how ordinary members of the public *evaluate scientific arguments* – that is, how they process arguments about scientific topics from a psychological perspective. How do people evaluate the sort of short and summarised scientific arguments (such as the IPCC quote, above) that they are exposed to in the media? Although several attempts have been made at studying people's ability to construct, deploy and evaluate scientific arguments (Driver, Newton & Osborne, 2000; Norris, Phillips & Korpan, 2003; Sadler, 2004; Simon, Erduan & Osborne, 2002; von Aufschaiter, Erduran, Osborne & Simon, 2008), the evidence pertaining to the public evaluation of scientific arguments is somewhat disparate, has mostly been conducted in educational settings, and is difficult to systematically piece together.

The reason for the lack of an account of how people interpret and evaluate scientific arguments is deceptively simple: A systematic framework for asking questions about how people evaluate scientific arguments has not been developed. In this chapter, I propose that the Bayesian approach to informal argumentation could be just such a framework.

The Bayesian approach has two distinct, but related advantages for studying scientific argument evaluation. Firstly, because the approach is *content-based*, it allows judgements about scientific arguments to be compared to judgements about non-scientific arguments. On the Bayesian account of argument strength, scientific arguments are simply arguments that *happen* to be about science. As such, they can be analysed in exactly the same way as non-scientific arguments. The key components that intuitively, might be thought to determine the strength of an argument (e.g. how much evidence it contains, the relation of the evidence to the hypothesis, the reliability of the source reporting the evidence) have a straightforward Bayesian interpretation. And, crucially, these factors can be identified and manipulated in both scientific and non-scientific arguments (Hahn & Oaksford, 2007a).

Secondly, because the approach is *normative*, it allows predictions to be made about when (and why) a particular argument should be strong or weak, depending on how much evidence it contains, or the reliability of the source of the evidence. As was discussed in Chapter 2, this approach (of identifying norms and comparing people's behaviour in experiments to them) has proven to be immensely popular and productive tool for organising research into human reasoning (see, e.g., Evans & Over, 1996; Nickerson, 2007; Oaksford & Chater, 1998). There is every reason therefore to expect it to be a productive way of studying informal scientific reasoning.

By bringing the Bayesian approach to bear on a number of different types of scientific arguments, a framework for examining scientific arguments can start to be developed that allows systematic questions to be posed about the factors that influence their strength. Are people bad at evaluating scientific arguments? Is there anything unique

or different about how people evaluate them? The answers to these questions have important implications for anyone interested in 'improving' science communication – if there is nothing special about scientific argument evaluation then there is no need to treat them as something distinct from other, non-scientific arguments. Consequently, education programmes aimed at improving science communication might benefit from existing knowledge about the evaluation of arguments in *general*. If, however, there are features unique to the evaluation of scientific arguments, then these features should be the focus of future research.

3.3   Science In Public – An Overview

The study of science communication is characterised by a multiplicity of approaches. There is a substantial philosophical literature on science as an epistemology (Knowles, 2003; Popper, 1959), as well as several prominent (and competing) sociological accounts of how science fits into the world of social actors, and how controversy and consensus develops in science (Brante, Fuller & Lynch, 1993; Collins & Pinch, 1993; Irwin & Wynn, 1996). How scientists construct bodies of knowledge, how these bodies of knowledge change, and what impact these changes have on research traditions has been a particular favourite of sociologists of science since T.S. Kuhn's (1970) influential work on the structure of scientific revolutions.

Most non-scientists' experience of evaluating science is likely to be through the media, and media analysts have sought to develop accounts of how different groups are involved in the production, communication and consumption of science (Friedman, Dunwoody & Rogers, 1999). Closely linked to these theories of

communication are attempts by social psychologists to understand how messages are effectively communicated – that is, the development of theories of persuasion (e.g., Chaiken, 1980; Petty & Cacioppo, 1984). Studying how people perceive risk and risky probabilities is another important component of understanding how science is perceived by the general public as typically, the communication of scientific information involves the communication of risk (see Pidgeon, Kasperson & Slovic, 2003, for a summary of work in this field).

Finally, there have been many attempts to measure people's attitudes and perceptions of particular scientific developments such as nuclear power (Bickerstaff, Lorenzoni, Pidgeon, Poortinga & Simmons, in press), climate change (Lorenzoni & Pidgeon, 2006) or nanotechnology (Pidgeon & Rogers-Hayden, 2007), typically utilising questionnaires and surveys (Poortinga & Pidgeon, 2003).

Even this cursory examination of some of the ways in which science communication has been approached highlights a fundamental problem facing researchers interested in the topic: It is not at all obvious where to start or which questions to ask. Clearly, in order to answer the question 'how do people evaluate science?' the wisdom of many different disciplines must be brought to bear. In order to understand science communication, however, it is essential to understand how non-experts *evaluate scientific arguments*.

3.4 Current Approaches to Understanding Scientific Argument Evaluation

Most of the existing research on scientific arguments has been qualitative (Sadler, 2004), emphasising people's *interpretation*, rather than their *evaluation* of arguments. Much of it has been focussed on tracking the development and quality of scientific reasoning (i.e. the use of hypotheses and evidence) in children (see, e.g., Klaczynski, 2000; Kuhn, Cheney & Weinstock, 2001; Kuhn & Udell, 2003). This literature is closely linked to educational policy and programmes designed to address deficits in scientific literacy, and provides a useful starting point from which to develop a more systematic framework for assessing how people evaluate scientific arguments (for a recent overview, see Sadler, 2004).

One approach to studying the use of scientific arguments has been to develop a qualitative taxonomy, or typology with which to analyse them. For example, Korpan, Bisanz, Bisanz and Henderson (1997) studied university students' evaluations of scientific news briefs. Responses were classified according to a hierarchical taxonomy, whereby questions relating to evidence, theory, social context and methods were deemed as indicating a high level of understanding, and therefore a more comprehensive evaluative approach. While most of the participants asked 'high level' evaluative questions on at least one occasion (e.g. requesting information about the statistics used, or the methodology employed), there were fewer questions about the social context of the scientific research (e.g. possible biases of people involved with the research, funding bodies etc)[2].

---

[2] Korpan et al also took measures of argument plausibility. Four topics of argument were used, three of which were designed to be scientific and 'plausible', and one (about a paranormal event) which was designed to be 'implausible'. Plausibility ratings were high for the three scientific topics, and low for the paranormal topic. Unfortunately, the plausibility and the topic (i.e. scientific or non-scientific) were

Using a similar classification system, Adams (2001) examined how college students, scientists and policy analysts evaluated 'questionable' scientific claims on the Internet about global warming. The qualitative responses they gave indicated that despite being less knowledgeable about the topics of the argumentation, the college students seemed able to apply 'generic' evaluative criteria to the reports, questioning the sources' validity, and disputing the degree to which it was appropriate to generalise or extrapolate given the available evidence. As participants were only given scientific claims to evaluate, however, it was not possible to compare the evaluation of scientific and non-scientific arguments – something that seems essential if claims about the *quality* of scientific argument evaluation are to be made.

Phillips and Norris (1999) focussed on evaluations of science reports from a popular science magazine, a non-science magazine or a newspaper. In this way, some attempt was made to contrast different types of argument – although in each case, the topic was scientific. In addition, the style of report was not manipulated systematically in the study, so no comparative data is available as to what effect each type of argument had. An attempt was made to measure *belief change*, albeit qualitatively, as participants were asked to indicate their background beliefs about each topic, and then report whether these beliefs had been altered by reading the report. The authors reported that participants' views became more polarised after reading the reports, but that the degree of certainty expressed by participants after reading the texts was inflated. Without some prediction about how convinced someone *should* be after reading a report, or how this new information should be integrated with their existing

confounded in this study, such that there was no way of comparing plausible vs. implausible scientific arguments – the stronger arguments were both scientific *and* plausible. However, this does at least give a preliminary indication that the evaluation of scientific arguments might not be wildly different to the evaluation of non-scientific arguments – a theme that I pursue in this Chapter.

beliefs, however, it is difficult to know what 'over certainty' or 'polarisation' might mean in this context. One advantage of the Bayesian approach is that the perceived strength of scientific arguments can be measured against a normative benchmark, and compared to non-scientific arguments – permitting precisely these sorts of judgements (i.e. whether a particular belief is 'polarised') to be made.

One particularly popular strategy in analysing student use and recognition of different types of scientific argument has been to apply the model of argumentation developed by Toulmin (1958). Toulmin's model is *dialectical*, in that it defines arguments as moves that are made in a conversation[3]. Because much of the scientific argumentation in educational settings occurs in a dialogue between the student and the teacher, several studies have drawn on Toulmin's model to develop an account of students' use of scientific arguments. According to Toulmin, an argument can be broken down into distinct components – a *claim* (the conclusion whose merits are to be established); *data* or *grounds* (the facts that are used to support the claim); *warrants* (the reasons that are used to justify the connections between the data and the claim) and *backing* (the basic assumptions that provide the justification for particular warrants). In addition, *qualifiers* (specifications of limits, or conditionality) and *rebuttals* (exceptions and counter-arguments to the claim) may be present. A schematic representation of Toulmin's model of argumentation is shown below in Figure 3.1.

---

[3] Van Eemeren & Grootendorst's pragma-dialectical theory (2004), which I discussed in detail in Chapter's 1 and 2, draws heavily on Toulmin's argumentation model. In the literature on scientific arguments, however, Toulmin's model, rather than the more recent pragma-dialectical approach, is typically used.
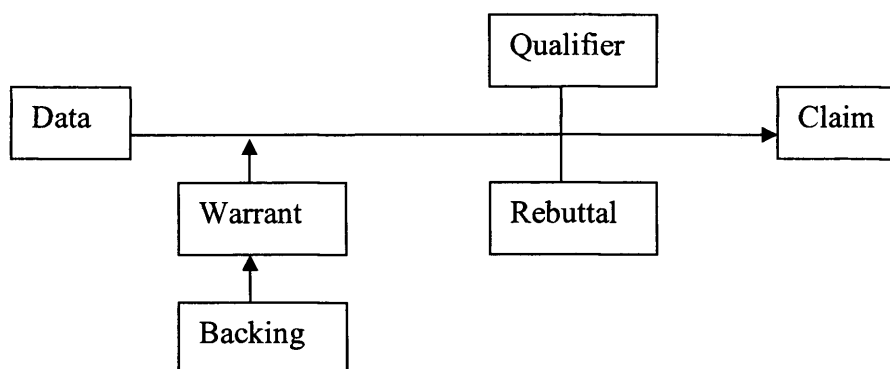
Figure 3.1: Schematic diagram of the Toulmin model of argumentation (Toulmin, 1958)

Because Toulmin's model describes the constitutive elements of an argument, and provides a method for comparing the structure of different arguments, it can be used to systematically analyse scientific argument *use*. Driver, Newton and Osborne (2000; see also Erduan, Simon & Osborne, 2004), for example, proposed that the model can successfully be used as a template for science educators to gauge the level at which their pupils are engaging with the issues being taught, by noting which argument features (e.g. warrants, data) are regularly being used in classes. According to Driver, Newton and Osborne, the ability to use (and recognise in other people) these different aspects of argumentation is an indicator of rhetorical competence, and therefore of comprehension (because understanding a subject is a prerequisite for reasoning competently about it). Drawing on these assumptions, Simon, Erduan and Osborne (2002) conducted an evaluation of the use of scientific argumentation in school science lessons, observing that:

"One of the many problems that bedevils work in this field is a reliable systematic methodology for a) identifying arguments and b) assessing quality. Our adoption and use of Toulmin has also provided us with a method for recognizing the salient features of argumentation..." (Simon, Erduran & Osborne, 2002, p22).

Using this approach, the authors analysed transcripts of teachers and students in science lessons, before and after the teachers had attended training courses aimed at increasing their understanding of different aspects of argumentation. Both students and teachers use of more complex forms of argumentation increased following the training courses.

However, whilst using Toulmin as a model of argumentation has been a popular tactic (Jiminez-Aleixandre, 2002; Jiminez-Aleixandre, Rodriguez & Duschl, 2000; Kortland, 1996), its application in individual studies has been inconsistent. Newton (1999), for example, observed the extent to which teachers provided opportunities for pupils to contribute to the co-construction of knowledge through discussion and argument. The types of activities that the pupils engaged in were then classified (e.g. listening, reading, question-answer interactions), but not directly related back to Toulmin's system of argument classification.

Von Aufschaiter, Erduran, Osborne and Simon (2008) used Toulmin's model to analyse the verbal conversations of school pupils during science lessons they had taken part in. Although Toulmin's model is primarily a system for classification, not *evaluation*, the authors identified patterns of argumentation that contained warrants and backings as demonstrating a higher quality of argument than argumentation based

on simply an unwarranted claim. Furthermore, if pupils were able to construct arguments with claims, warrants and data even once their initial position had been rebutted, they were classified as demonstrating an even higher level of argument quality. While this approach offers a systematic way of comparing the pupils' *use* of arguments, however, it is not necessarily informative about their *quality*. As Driver, Newton and Osborne (2000) point out, this is because;

"Toulmin's analysis...is limited as, although it can be used to assess the structure of arguments, it does not lead to judgements about their correctness...it is necessary, if judgements of this kind are to be made, that subject knowledge is incorporated for arguments to be evaluated" (Driver, Newton & Osborne, 2000, p294).

In other words, two arguments can have an identical logical form, be given in exactly the same dialectical circumstances, but still differ substantially in their convincingness – because they differ in their *content*. This issue is, of course, not limited to Toulmin's model – these are precisely the same criticisms that I raised earlier (see Chapter 1 and Chapter 2) in relation to procedural theories in general – and to van Eemeren and Grootendorst's (2004) pragma-dialectical theory specifically.

Toulmin's model is a powerful tool for specifying the component parts of arguments, and classifying them accordingly. It deals only, however, with the structure of different arguments, not their content. Hence, it is not a measure of argument strength. Von Aufschaiter et al (2008) circumnavigated this problem by developing a system of content analysis that they used in parallel with Toulmin to assess argument quality. They measured the learning that occurred following argumentation and proposed that

an increase in learning indicated more sophisticated argumentation. Ideally, however, an account of scientific argument strength would be sensitive to argument content, and would not necessitate the development of a separate system of content analysis – or as von Aufschaiter et al (2008) put it "a more careful consideration of the interrelationship between the content *and* process of an argument" (von Aufschaiter et al, 2008, p128).

There have been other investigations into how people evaluate scientific arguments (Kolsto, Bungum, Arneses, Isnes, Kristensen, Mathiassen, Mestad, Quale, Tonning & Ulvik, 2006; Kuhn, Shaw & Felton, 1997; Norris, Phillips & Korpan, 2003; Patronis, Potari & Spiliotopolou, 1999; Ratcliffe, 1999; Takao & Kelly, 2003). These have tended to employ qualitative classification systems developed on an ad hoc basis (Takao & Kelly, 2003). They therefore provide little opportunity for comparison with other studies, and frequently apply only to a specific set of data (Kuhn, Shaw & Felton, 1997). As Kolsto et al (2006) note about their own classification system;

"It is important to be aware that the identification of these criteria was done in a specific context. If the context had not been educational, if the students had selected other articles, and if students with other experiences and knowledge had been involved, other or additional criteria might have been found...Also, we have not made any systematic normative analysis..." (Kolsto et al, 2006, p646).

This conclusion is not the strongest basis for elucidating the general factors that influence the evaluation of scientific arguments. It should be emphasised, of course, that the vast majority of the studies considered above were not explicitly aimed at

developing an account of argument strength – rather, they were primarily concerned with *improving* scientific literacy in educational settings. Their consideration allows the identification, however, of the important themes that arise and suggests that the barriers to developing a more systematic account of scientific argument evaluation are twofold.

Firstly, evaluative assessments of argumentation are made either using a framework developed specifically for the study or using Toulmin's (1958) model. While the use of Toulmin's approach to argument classification has been enormously beneficial to science educators wishing to improve scientific literacy in educational contexts, it has some significant limitations as a general purpose framework for understanding the evaluation of scientific arguments – primarily because it focuses on argument *form*, rather than argument *content*[4]. What this means is that while the different components of argumentative discourse can be identified and classified (and interventions designed to increase the sophistication of the argumentation) Toulmin must be supplemented with an analysis of content to act as a measure of argument *quality* (see von Aufschaiter et al, 2008). Analysing arguments by their content also means that it is straightforward to measure their strength quantitatively – as judgements are not limited to the qualitative classification of argument components. Quantitative measures allow more detailed and specific predictions to be made.

Secondly, if an assessment of how *good* people are at evaluating scientific arguments is to be made, then a normative approach is required. Studies explicitly aimed at comparing the evaluation of scientific and non-scientific arguments are scarce, yet

---

[4] I have made similar criticisms of procedural theories of argumentation in general, in Chapter 1 and Chapter 2.

this seems essential if we are to make claims about the relative strength of scientific and non-scientific arguments, and people's capacity to evaluate them.

What is clear from considering the existing literature on scientific argument evaluation is that an approach that permitted quantitative predictions about argument strength to be made, key variables to be experimentally manipulated, and performance across a number of tasks to be meaningfully compared would be invaluable in improving our understanding of the evaluation of scientific arguments. At a minimum, it seems necessary to have an approach that allows the systematic comparison of scientific and non-scientific argument evaluation, a putative normative framework for formulating questions about argument strength, and measures that are sensitive to changes in argument content as well as form and context. In the following section, I will suggest that the Bayesian framework for informal argumentation may be just such an approach.

3.5   The Bayesian approach to scientific argument evaluation

There is a growing body of evidence that the Bayesian approach provides a suitable framework for analysing and investigating different types of informal argument (Hahn & Oaksford, 2007a; Hahn, Oaksford & Corner, 2005; Oaksford & Hahn, 2004). From the Bayesian perspective, scientific arguments are simply arguments that *happen* to be about science and can be analysed in exactly the same way – a strong scientific argument is simply one that provides evidence in support of a claim.

For example, Oaksford and Hahn (2004) investigated a type of argument known as the 'argument from ignorance'. These are simply arguments whereby the absence of

evidence (e.g. finding *no* side effects when testing a new drug) is used to support a hypothesis (e.g. that the drug is safe). Oaksford and Hahn used a Bayesian approach to try and capture some of the quantities that (intuitively) make these arguments seem more or less compelling, and then made predictions about how favourably the arguments would be evaluated by participants in an experiment. I will describe the experiment in some detail, as it has some important implications for the application of a Bayesian framework to the analysis of scientific arguments.

Oaksford and Hahn identified several components that should, from a Bayesian perspective, affect the strength of an argument from ignorance, which are best illustrated with an example (also discussed in Walton, 1992a). Imagine that someone were to claim that a train does *not* stop at a certain station, because the station is not listed on the timetable. This is an argument from ignorance that seems perfectly plausible. The argument is comprised of a claim (that the train does not stop at a certain station) and the negative results of a test (consultation of the timetable). These negative test results are the evidence on which the claim is based. Oaksford and Hahn suggested that the convincingness of this sort of argument will depend on particular *characteristics of the test*, which can be given a very simple Bayesian interpretation. A train timetable, as a test of which stations the train will stop at, is a highly *sensitive* test, because it has a high 'hit rate' $P(e \mid h)$. That is, the train will stop at *all* the stations on the timetable. It is also a highly *selective* test – as the train will not stop at any stations *not* listed on the timetable. This means that its false-positive rate, $P(e \mid \neg h)$ is low. By contrast, a shop window is not necessarily a particularly sensitive or selective source of information – just because a particular piece of clothing does not appear in the window, it seems unjustified to conclude that the shop

does not stock it. These test characteristics, which have a direct influence on how strong the evidence is that it provides, feed directly into our evaluation of how strong the argument from ignorance is. When the evidence is highly diagnostic (i.e. a high hit rate and a low false positive rate), the argument should be stronger (see Chapter 1 for a more detailed discussion of the Bayesian treatment of evidence). Oaksford and Hahn found that arguments from ignorance that contained more diagnostic evidence were rated as more convincing across a range of topics.

By conceptualising judgments of argument strength as subjective degrees of belief, quantitative measurements can be taken and predictions made about the relative strengths of different types of arguments. The factors that seem important in the evaluation of scientific arguments, such as the amount of evidence an argument contains, or the reliability of the source providing the evidence, can be simply and systematically represented using the Bayesian framework. And because the Bayesian approach is a general account of argument strength comparisons between different experimental tasks and between scientific and non-scientific argument evaluations are straightforward. All the evidence thus far (Hahn & Oaksford, 2007a; Hahn, Oaksford & Corner, 2005; Oaksford & Hahn, 2004) suggests that people are perfectly good at evaluating factors such as evidence and source reliability in non-scientific arguments. On the Bayesian account it is the very same quantities that determine the strength of informal scientific arguments.

The goal of much of the science communication literature is, however, to *improve* the understanding of (and reasoning about) scientific issues (Irwin & Wynne, 1996; von Aufschaiter et al, 2008). This would suggest that there might be something unique

about scientific arguments. Are they weaker, or less effective than non-scientific arguments? Do different factors influence their strength? If there are discrepancies between scientific and non-scientific argument evaluation, then these discrepancies have important implications for the 'improvement' of science communication: In order to understand science communication, we must first have an adequate theory of informal argument strength.

In the remainder of this chapter, I will report the results of four experimental tasks designed to test some basic and straightforward predictions that fall out of the Bayesian approach. My aim is to identify a) whether people evaluate different types of scientific and non-scientific arguments in line with Bayesian predictions; and b) whether people evaluate scientific arguments any differently from non-scientific arguments.

3.6 General Methods and Participant Information

Four experiments were designed, each focusing on a different aspect of people's evaluation of scientific arguments or evidence. The same sample of one hundred participants completed three of the experiments, as part of a single experimental session. A separate sample of one hundred participants took part in the remaining experiment. Both samples were comprised of undergraduate students from the Psychology Department at Cardiff University, who participated in the experimental sessions in exchange for course credit. I have labelled the experiments 1a – 1d, as the experiments all relate to different aspects of the evaluation of scientific arguments. The experiment that was conducted with the second sample appears here as 1c.

Experiments 1a, 1b and 1d were conducted in the first session. Although experiments with separate samples of participants are conventionally labelled separately, this method of presentation was the clearest way of reporting the results of the four experiments.

Participants within each session were assigned the same tasks, but occasionally a task was not completed by a participant. The number of participants that complete each experimental task is reported in the results section for each experiment. The order that the first sample of participants completed the three experiments was counter-balanced – participants who took part in the second session completed one task only. The gender of participants was not recorded for any of the experiments reported in this thesis.

3.7   Experiment 1a – Arguments from Ignorance

The first experiment is based on the work of Oaksford and Hahn (2004), discussed in detail above. Whilst most arguments are based on the *observation* of evidence, some arguments are based on the *absence* of evidence. These so-called 'arguments from ignorance' use the absence of evidence (e.g. no side effects found during the testing of a new drug) to support a hypothesis (e.g. that the drug is safe). This is a particularly common type of argument found in scientific discourse, as the familiarity of the above example demonstrates. Pharmaceutical companies regularly employ them to assure us of a new product's safety – in fact the industry safety standard involves demonstrating 'no harmful side effects' over a given testing period. We seem happy to accept these arguments (as evidenced by our willingness to take the drugs), and assume that we

can be confident in them because the search for evidence has been thorough, and the type of test conducted appropriate. In fact, many high-profile socio-scientific arguments seem to take the form of an argument from ignorance – arguments about the safety of nuclear power stations or the Measles, Mumps and Rubella (MMR) jab are both founded on a *lack* of evidence that any danger exists. Similarly, debates about health epidemics such as Creutzfeldt-Jakob Disease (CJD) and Aviation Influenza ('Bird Flu') are often based on the *absence* of evidence that an outbreak will occur.

Four arguments from ignorance were developed – two based on scientific topics, and two based on non-scientific topics. In an attempt to capture the key quantities that determine the strength of arguments from ignorance, the credibility of the source giving the argument (either a reliable or an unreliable source) and the type of search they conducted for evidence (either a thorough, or an incomplete search) was manipulated. Thus, as the diagnosticity of the evidence increases, so should ratings of argument strength. Consider the following two examples of arguments from ignorance:

Dave: This new anti-inflammatory drug is safe.
Jimmy: How do you know?
Dave: Because I read that there has been one experiment conducted, and it didn't find any side effects.
Jimmy: Where did you read that?
Dave: I got sent a circular email from exciting news@wowee.com

Dave: This new anti-inflammatory drug is safe.

Jimmy: How do you know?

Dave: Because I read that there have been fifty experiments conducted, and they didn't find any side effects.

Jimmy: Where did you read that?

Dave: I read it in the journal Science just yesterday.

Intuitively, the second argument seems more compelling – the search for evidence is more thorough, and the (absence of) evidence is reported by a more reliable source. Figure 3.2 shows that this intuition is captured well by the process of Bayesian updating.

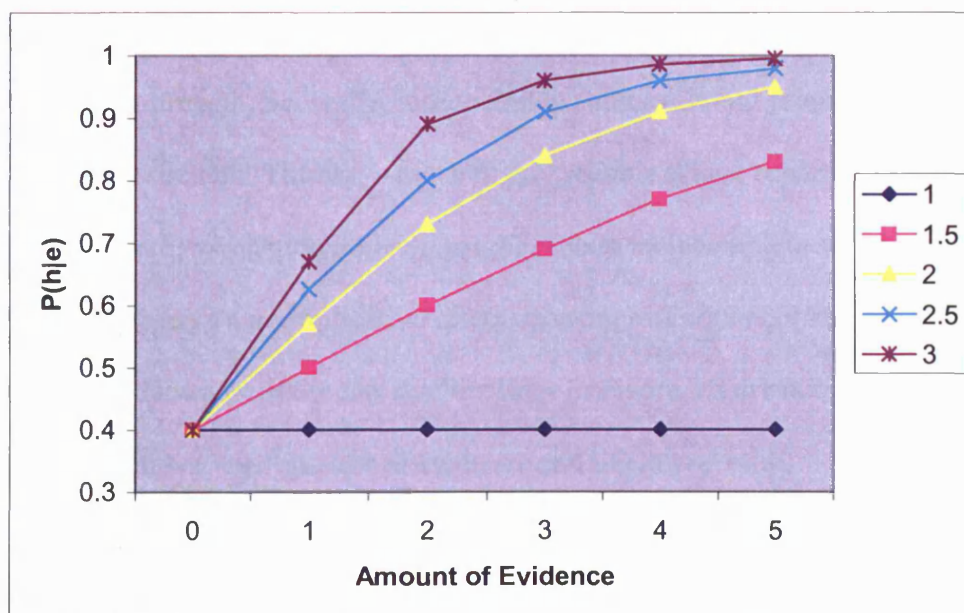*Figure 3.2*: Impact of amount of evidence and source reliability (likelihood ratio) on posterior belief in a hypothesis. Each line represents a different likelihood ratio.

The graph plots the impact of increasing 'units' of evidence (e.g. the results of experiments that find no side effects for a drug), and an increasing likelihood ratio (i.e. an increasingly reliable source reporting these results) on posterior degree of

79

belief in a hypothesis ($P(h \mid e)$). Starting from a prior degree of belief of 0.4, a posterior degree of belief is calculated following the addition of each 'unit' of evidence. This posterior then becomes the new prior for the next 'unit' of evidence, and updating proceeds from there. Where the likelihood ratio is low; that is, where a source of evidence is just as likely to report $P(e \mid h)$, as they are to report $P(e \mid -h)$, the impact of the evidence has little impact. When the likelihood ratio is higher, however, increasing the amount of evidence has a systematic effect on posterior belief in the hypothesis.

From Figure 3.2 it is possible to make three predictions. Firstly, increasing the amount of evidence (i.e. a more thorough search for evidence) should produce higher ratings of argument strength. Secondly, more reliable sources should produce higher ratings of argument strength. Thirdly, when a highly reliable source reports a thorough search for evidence (or when an unreliable source reports an incomplete search for evidence) we should observe a multiplicative effect on ratings of argument strength. This third prediction follows from the fact that the lines in Figure 3.2 are not parallel – there is an interaction between amount of evidence and likelihood ratio.

Additionally, I will compare the evaluation of scientific and non-scientific arguments from ignorance. Given that arguments from ignorance are so prevalent in popular scientific discourse, one might expect acceptance of arguments from ignorance to be higher in scientific domains than in non-scientific domains. To the best of my knowledge this experiment is the first comparison of scientific and non-scientific arguments from ignorance, and so strong predictions would be premature. It seems likely, however, that the manipulation of two key variables in a type of argument

frequently used in scientific discourse will provide a clearer picture of how these variables influence judgements of argument strength – something which has not been possible in previous investigations into scientific argument evaluation.

3.7.1 Methods

Design

Four arguments based on the absence of evidence were developed. Three variables were manipulated (Source, Search and Class) at two levels (reliable/unreliable, thorough/incomplete and scientific/non scientific) across four different argument topics, creating a total of sixteen distinct arguments. The topic, type and order of the arguments were randomised using the Latin Square Confounded method where participants see only one argument from each topic, and participate once in each experimental condition (Kirk, 1995). This allows multiple responses to be obtained from each participant, but prevents arguments about the same topic being viewed by any one participant.

The Latin Square Confounded design requires that the topics are rotated across the four conditions of the Search X Source manipulation. This means that although all participants see one argument from each topic, and one argument from each level of the Search X Source manipulation, the combination of the variables differs systematically across participants. So, for example, one participant might receive the following set of arguments:

S1 (Thorough/Reliable) – **S2** (Thorough/Unreliable) – **NS1** (Incomplete/Reliable) – **NS2** (Incomplete/Unreliable)

A second participant might receive the following set of arguments:

S1 (Thorough/Unreliable) – S2 (Thorough/Reliable) – NS1 (Incomplete/Unreliable) – NS2 (Incomplete/Reliable)

Because each argument topic can embody one of four experimental conditions (the four levels of the Search X Source manipulation), sixteen distinct arguments are created. And because participants must only see one of each topic and one argument from each experimental condition, sixteen distinct combinations are possible. The Latin Square Confounded design simply ensures that each of the sixteen combinations are represented an equal number of times in the sample (or as close to equal as is possible given the sample size). In addition, the *order* that the arguments are received within each combination is randomised. The Latin Square is therefore an effective way of systematically varying topics and experimental conditions, while guarding against order effects.

Participants were required to indicate how convincing they found the arguments on a scale from 0 (very unconvincing) – 10 (very convincing), and also how reliable they thought the source in the argument was on a scale from 0 (unreliable) – 10 (very reliable).

Materials and Procedure

Each participant received an experimental booklet containing four arguments from ignorance on different topics. The two science topics were: (S1) the safety of a new anti-inflammatory drug, and (S2) the risks associated with GM crops, and the two non-scientific topics were: (NS1) the release of a new games console, and (NS2) the

presence or absence of a particular item of clothing in a High Street store. The four

topics are shown below in Figure 3.3. Information pertaining to the experimental

manipulation of the Search and Source variable is highlighted in bold.

S1

Dave: This new anti-inflammatory drug is safe.

Jimmy: How do you know?

Dave: Because I read that there **has been one experiment conducted/have been fifty**

experiments conducted, and **it/they** didn't find any side effects.

Jimmy: Where did you read that?

Dave: **I read it in the journal Science just yesterday/I got sent it from**

excitingnews@wowee.com.

S2

Gemma: GM crops are safe to eat.

Kate: How do you know?

Gemma: Because I read that there has been **one experiment/fifty experiments** conducted,

and **it/they** didn't find any side effects.

Kate: Where did you read that?

Gemma: I think I read it **in the New Scientist/on someone's internet blog or something**

NS1

Aaron: The new upgraded Playstation console hasn't been released yet.

Celia: How do you know?

Aaron: Because I checked **one website/fifty websites** and they didn't have it in stock

Celia: Where did you look on the internet?

Aaron: I used a price comparison search engine called **www.gamesconsoles.com/**

www.KitchensandHomeware.com.

NS2

Alicia: There are no red dresses left in any TopShop store in the UK

Greg: How do you know that?

Alicia: I phoned **one store/fifty stores** to see if they had any left

Greg: How did they check to see if they had any in stock?

Alicia: **The store manager checked on the shop's stock database/they had a quick look while I was on the phone.**

Figure 3.3: An example of each topic used in Experiment 1a. Information pertaining to the experimental manipulation of the Search and Source variables is highlighted in bold.

3.7.2   Results and Discussion

99 participants completed the experiment. Every participant saw four arguments – one from each experimental condition, and one of each topic according to the Latin Square Confounded experimental design described above. To statistically analyse data from Latin Square Confounded designs, participant effects within the ratings are factored out and the analyses are conducted on the residuals (Kirk, 1995)[5].

Figure 3.4 displays the mean ratings of argument strength obtained in each experimental condition. Although residual ratings were used for statistical analyses, all graphed data is raw. The residual transformation tends to cluster responses, making

---

[5] Computing residual values is necessary because although participants provide data in every condition of the experiment, the combination of topic and experimental condition differs between participants. Computing a residual transformation permits standard, between-subjects analyses to be conducted. Though this changes the absolute numerical values, it typically leaves the overall shape of the data unaltered. In all the data reported in this thesis, analyses of variance on raw and residual values produced the same statistical effects.

the visual interpretation of residual data difficult. Displaying raw, rather than residual

data, permits a more natural interpretation of participant responses.
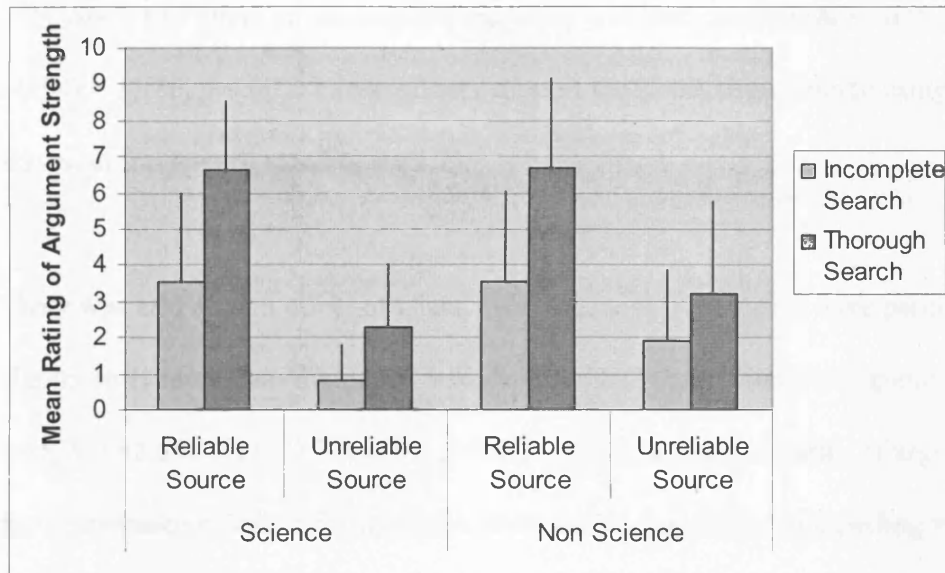
Figure 3.4: Mean ratings of argument strength across each condition of the experiment. Error bars indicate one standard deviation.

A three-way ANOVA was conducted on the residual ratings of argument strength, with Search (thorough/incomplete), Source (reliable/unreliable) and Class (scientific/non-scientific) as independent variables. The presence of a thorough search for evidence in the argument produced higher ratings of argument strength ($M = 1.14$, $SD = 2.56$) than an incomplete search ($M = -1.13$, $SD = 2.06$), $F (1, 387) = 156.29$, $p < .001$. A reliable source reporting the results of the search produced higher ratings of argument strength ($M = 1.4$, $SD = 2.34$) than an unreliable source ($M = -1.41$, $SD = 1.97$), $F (1, 387) = 269.57$, $p < .001$. Furthermore, there was a significant interaction between Search and Source, $F (1, 387) = 20.19$, $p < .001$. Pairwise comparisons showed that for unreliable sources, a thorough search for evidence produced significantly higher ratings of argument strength ($M = -0.65$, $SD = 1.99$) than an

incomplete search for evidence (M = -2.2, SD = 1.06), t (194) = 6.02, p <.001. For

reliable sources, however, this difference was much greater. A report of a thorough

search by a reliable source produced far higher ratings of argument strength (M =

2.93, SD = 1.63) than an incomplete search by a reliable source (M = -0.9, SD = 1.93),

t (194) = 11.86, p <.001. These effects support the predictions I made using the

Bayesian framework (see Figure 3.2).

There was also a main effect of Class. Non-scientific arguments were perceived as

significantly more convincing (M = 0.15, SD = 2.5) than scientific arguments (M = -

0.15, SD =2.65), F (1, 387) = 2.05, p < .001. Despite the familiarity of arguments

from ignorance in scientific discourse, they were rated as less compelling than non-

scientific arguments. Reference to Figure 3.4, however, suggests that non-scientific

arguments were not *universally* preferred over scientific ones, and, in fact, the

interaction between Source and Class was also significant, F (1, 387) = 11.32, p

<.001. Pairwise comparisons indicated that participants' evaluations of scientific and

non-scientific arguments only differed under certain conditions. While there was no

significant difference between ratings of reliable scientific (M = 1.41, SD = 2.27) and

reliable non-scientific arguments (M = 1.39, SD = 2.44), unreliable scientific sources

produced significantly lower ratings of argument strength (M = -2.16, SD = 1.5) than

unreliable non-scientific sources (M = -0.8, SD = 2.01), t (194) = 4.99, p <.001. While

this effect was not predicted, the ratings of source reliability that participants provided

may go some way to explaining it.

Source reliability ratings are displayed in Figure 3.5. A three-way ANOVA was

conducted on the residual source reliability ratings. Judgements of source reliability

were significantly higher when the source was reliable ($M = 2.24$, $SD = 2.09$) than

when the source was unreliable ($M = -2.27$, $SD = 1.61$), $F (1, 387) = 675.03$, $p < .001$.

Sources were also judged to be more reliable when the test conducted for evidence

was thorough ($M = 0.6$, $SD = 2.89$) than when it was incomplete ($M = -0.6$, $SD =$

2.85), $F (1,387) = 49.22$, $p < .001$. Although the effect of Class was non-significant,

there was a significant interaction between Source and Class of argument, $F (1, 387)$

$= 25.97$, $p < .001$.
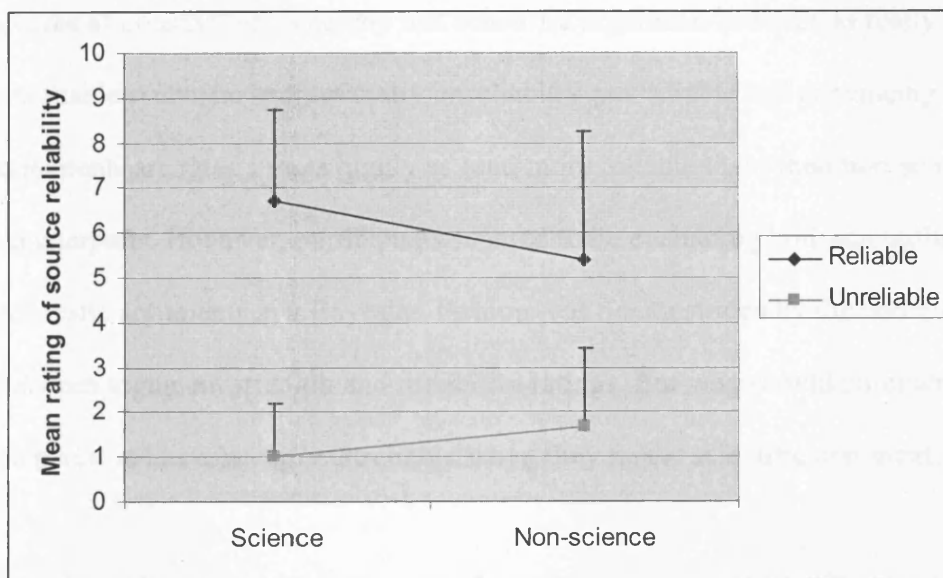
Figure 3.5: Mean ratings of source reliability obtained in each experimental

condition. Error bars indicate one standard deviation.

Pairwise comparisons showed that unreliable scientific sources were perceived as

significantly less reliable ($M = -2.72$, $SD = 1.25$) than unreliable non-scientific

sources ($M = -1.91$, $SD = 1.78$), $t (194) = 3.59$, $p < .001$, but also that reliable

scientific sources were significantly more reliable ($M = 2.7$, $SD = 1.73$) than reliable

non-scientific sources ($M = 1.66$, $SD = 2.36$), $t (194) = 3.56$, $p < .001$. This

polarisation of ratings of scientific source reliability tracks the pattern observed in the

convincingness ratings. Scientific arguments from ignorance from unreliable sources

may therefore have been rated as less *compelling* because participants judged these

sources to be less *reliable*. These reliability ratings provide a possible explanation for

the argument strength ratings: Less reliable sources should produce less compelling

arguments (and more reliable sources more compelling arguments) on the Bayesian

account. This is exactly the pattern of results seen in the argument strength ratings.

Taking account of the data from both the dependent measures, there seems to be a

degree of polarisation, whereby bad scientific arguments are seen as really bad, and

unreliable scientific sources really unreliable – yet reliable and convincing scientific

arguments are rated just as highly as (and more reliable than) their non-scientific

counterparts. However, participants seemed to be evaluating both scientific and non-

scientific arguments in a Bayesian fashion – as demonstrated by the correspondence

between argument strength and reliability ratings. But why should unreliable sources

be perceived as *especially* unreliable when they report scientific arguments?

One possible explanation for the polarised ratings of scientific argument strength and

reliability is the perceived position of scientific knowledge in our lives. Scientific

knowledge is taught in the classroom as a series of facts – certainties – that are arrived

at by a rigorous and objective process of hypothesis testing (Simon, Erduan &

Osborne, 2002). When this scientific method seems to be absent from a scientific

argument – as when an argument is offered from an unreliable source and appears to

be based on an incomplete search for evidence – we immediately doubt its validity. In

other domains, we might not have such strict criteria[6]. We are used to evaluating non-

---

[6] This is supported by the data in this experiment – unreliable sources are perceived as worse than
reliable sources for non-scientific arguments, but this effect is not as pronounced as it is in the
scientific topics.

scientific arguments that are not based on solid evidence, or where the source is less than fully reliable. Most of our non-scientific knowledge is not obtained by the scientific method. It is perhaps no surprise that a scientific argument lacking evidence from an unreliable source strikes us as particularly unconvincing, and equally plausible that a reliable and evidence based scientific argument is more compelling than an comparable non-scientific argument[7].

This explanation is necessarily speculative, and requires further investigation. But by quantifying and manipulating two key properties of scientific argument evaluation, and comparing the evaluation of scientific arguments from ignorance to non-scientific arguments from ignorance, it has been possible to give a fairly detailed account of how they are evaluated. While concepts such as source reliability (Kolsto, 2001), evidential strength (Korpan et al, 1997), and belief polarisation (Phillips & Norris, 1999) have been discussed in previous attempts to investigate scientific argument evaluation, applying a Bayesian framework has allowed the isolation and measurement of these variables – all the while using non-scientific arguments as a comparison point.

---

[7] There is one other aspect of the data that requires explanation – ratings of source reliability were consistently higher when the search for evidence was more thorough. Despite not impacting directly on the reliability of the source (i.e. a blog that reports fifty experiments is still a blog), the type of search that the source in each argument conducted had an effect on how reliable that source was perceived to be. It is possible, however, that the extent to which a source is able to conduct a valid search for evidence (even an inappropriate or unreliable source) bears on the subsequent evaluation of the source itself. In particular, it is worth considering *why* an unreliable source is perceived this way in the first place – one reason might be that the source has a reputation for providing weak evidence. If a previously unreliable source starts to report what appears to be stronger evidence for its claims, this is likely to impact on subsequent judgements of that source's reliability: There is a reciprocal relationship between the reliability of a source and the evidence it provides. The inflated ratings of source reliability in this experiment may simply be a function of this relationship.

The general pattern of data supports the results obtained in Oaksford and Hahn .(2004), and suggests that both scientific and non-scientific arguments from ignorance are evaluated in line with Bayesian predictions. Where differences emerged between scientific and non-scientific argument evaluation they were accounted for by differences in judgements of source reliability. I have offered a tentative explanation for why the science arguments displayed a degree of polarisation although future research would be required to establish whether this is a reliable phenomenon. The results of the first experiment suggest, however, that valuable insights into the process of scientific argument evaluation – and therefore science communication in general – can be made using simple experimental tasks, and with the application of a clear theoretical framework. In the next experiment, I examine the notion of mixed or contradictory evidence in the evaluation of hypotheses, using the Bayesian approach to pose questions about how people respond to evidential *uncertainty*.

3.8   Experiment 1b – Uncertainty and Mixed Evidence

The communication of uncertainty in scientific messages is the subject of much controversy – not least in current debates about climate change (Patt, 2007; Zehr, 2000). On the one hand, practitioners of science understand that the scientific method inevitably produces uncertainty. Some sciences in particular (and climate prediction is one of them) are inherently difficult to practise with precision. Trying to predict what will happen with any certainty in a complex system is an immensely complicated task. Any particular hypothesis (e.g. that anthropogenic climate change will cause an increase in the average global temperature of 2 degrees centigrade above pre-industrial levels within 50 years) is likely to be the product of many different

experiments (in this case, many different climate models), each of which is likely to have produced slightly different data (e.g. a range of predicted temperature increases of between 1 and 3 degrees centigrade within 50 years). In peer-reviewed journals, different laboratories publish data that may support or contradict existing findings. Does this indicate that the science in some of these studies must be misleading, or unreliable? It may do, of course, but observing a range of findings when attempting to answer a complicated scientific question is entirely normal, and generally, scientists themselves do not find this natural variation a cause to label a research programme flawed or inaccurate. Indeed, that any consistent patterns can be observed at all in sciences involving complex prediction suggests the existence of a strong effect. That the IPCC can claim that anthropogenic climate change is *very likely* to be occurring indicates significant scientific consensus (IPCC, 2007).

As Pollack (2005) notes;

"Far from being an impediment that stalls science, uncertainty is a stimulus that propels science forward" (Pollack, 2005, p5).

The use of different experimental and statistical techniques, the interpretation of ambiguous data, and chance factors will all influence the conclusions reached in scientific studies. It is only when a consistent pattern of reliable and replicable data emerges across different experimental programmes that something resembling a consensus emerges. But even then, with a pattern of data that the majority of experts would be willing to call a reliable conclusion, there will be deviations. A good

scientist should be suspicious of a scientific consensus where there are *no* exceptions – competing hypotheses are an essential part of scientific epistemology (Kuhn, 1970).

Of course, this insight into the scientific method is something that is for the most part confined to scientists themselves (although see Collins & Evans, 2007, for information on the types of scientific expertise that are possessed by different groups in society). What the general public receive are carefully packaged representations of science – typically communicated by the media, and never containing anything like enough detail to allow a 'complete' evaluation to be made. Instead the public consume science as a series of facts and hypotheses – and, analogously to the tuition of science at school as facts and certainties – come to evaluate it as either right or wrong. This reduction of what is a complex and inherently variable process into a 'sound bite' of scientific information inevitably leads to polarised views of scientific conclusions – and when competing hypotheses are reported by the press, they may appear bafflingly contradictory. The existence of *any* uncertainty about the impact of anthropogenic activity on the environment is often interpreted by the media as indicating a lack of scientific consensus. Without access to the 'nuts and bolts' of scientific research, the normal variation of the scientific method may be perceived as incoherence and confusion from the scientific community – which ultimately leads to distrust of the scientific data (Pollack, 2005).

There is a straightforward way in which the notion of uncertainty and contradictory reports can be translated into an experimental task using the Bayesian framework as a guide. In Bayesian terms, a good argument is one that gives evidence in support of a hypothesis. The presence of mixed or contradictory evidence should lead to a lower

degree of belief in this hypothesis than if the evidence all supported it. But having

only mixed evidence in support of a hypothesis is not the same as having evidence

that *disconfirms* a hypothesis – mixed evidence should provide an intermediate level

of support for a hypothesis. Because the evaluation of evidence and the extent to

which it supports a hypothesis is a crucial part of the Bayesian account of argument

strength, it is possible to use this framework to predict when belief in a hypothesis

should be high, when it should be low, and when it should be somewhere in-between.

While this may seem intuitive, it is only when an appropriate framework can be

brought to bear on problems such as these that predictions can be made at all – so

while it may be obvious that confirming evidence should be more compelling than

disconfirming evidence, the Bayesian framework provides a simple model with which

to test these intuitions (see Figure 3.2 for an indication of how increasing amounts of

evidence should impact on posterior degree of belief).

Recent Bayesian treatments of evidence evaluation have also suggested that

coherence is an important factor in how favourable a set of evidence is judged to be

(Bovens & Hartmann, 2003). In particular, evidence may impact negatively on the

truth of a hypothesis, but be consistent in doing so – that is, all the evidence is

negative, and as such, the evidence is *coherent*. Bovens and Hartmann (2003)

demonstrated that the coherence of an information set and the reliability of the

sources that provide this information both determine its convincingness. This means

that although we might expect that the *strength* of an argument should increase

according to the amount of confirmatory evidence it contains, the *reliability* of the

sources who provide the evidence are affected by the evidential coherence. Therefore

in addition to obtaining measures of argument strength, participants were also asked

to indicate how reliable they thought the sources of evidence were. A Bayesian analysis predicts that while mixed evidence should produce ratings of argument strength that are intermediate, incoherent evidence should produce ratings of source reliability that are *lower* than consistently disconfirming evidence.

Finally, while practising scientists may be aware that uncertainty is an integral part of the scientific process, the public perception of science as a discipline that provides certainty and consistency does not sit comfortably with the notion of incoherent scientific evidence. In Experiment 1a, ratings of scientific arguments from ignorance were polarised. The degree to which evidential coherence impacts on ratings of source reliability may therefore be greater for arguments about scientific topics.

3.8.1   Methods

Design

Two arguments were developed (one scientific and one non-scientific) based on a claim and some evidence. Four pieces of evidence accompanied each claim, from four different sources. Two variables were manipulated in this experiment; Evidence (confirms/mixed/disconfirms) and Class (scientific/non-scientific). This created a total of 6 distinct arguments. Participants evaluated one argument from each topic (evaluating two arguments in total), and the type and order of the arguments were randomised using the Latin Square Confounded design described in Experiment 1a. Participants were required to indicate how likely they thought the claims in the arguments were to be true on a scale from 0 (Unlikely) – 10 (Very Likely), and also to

indicate how reliable they thought the sources providing the evidence were, on a scale from 0 (Unreliable) – 10 (Very Reliable).

Materials and Procedure

Each participant received an experimental booklet containing two claim-plus-evidence arguments on different topics. The science argument concerned the date of the next visible solar flare, and the non-scientific argument concerned the date of the leader of the Conservative Party and his wife's expected baby. Figure 3.6 shows the claim and the three levels of the evidence manipulation for both arguments.

SCIENCE CLAIM:

The next visible solar flare will be during the month of November.

CONFIRMING EVIDENCE:

Professor Grantham has calculated that the next visible solar flare will occur on November 20th.
Professor Bootley reports that there will be a visible solar flare in the first week of November.
Professor Parry has identified November as the most likely date for the next visible solar flare
Professor Reddon published a statement which estimated the next visible solar flare to occur during the winter.

MIXED EVIDENCE:

Professor Grantham has calculated that the next visible solar flare will occur on November 20th.
Professor Bootley reports that there will be a visible solar flare in the first week of August.
Professor Parry has identified November as the most likely date for the next visible solar flare
Professor Reddon published a statement which estimated the next visible solar flare to occur during the summer

DISCONFIMING EVIDENCE:

Professor Grantham has calculated that the next visible solar flare will occur on July 20th.
Professor Bootley reports that there will be a visible solar flare in the first week of August.
Professor Parry has identified July as the most likely date for the next visible solar flare
Professor Reddon published a statement which estimated the next visible solar flare to occur during the summer.

NON-SCIENCE CLAIM

The leader of the Conservative Party and his wife are going to have a baby in the summer.

CONFIRMING EVIDENCE:

The Daily News reports that the leader of the Conservative Party and his wife are expecting a baby in August.

The Globe reports that the leader of the Conservative Party and his wife are expecting a baby in July.

The World Today reports that the leader of the Conservative Party and his wife are expecting a baby in June.

News Update reports that the leader of the Conservative Party and his wife are expecting a baby in May.

MIXED EVIDENCE:

The Daily News reports that the leader of the Conservative Party and his wife are expecting a baby in November.

The Globe reports that the leader of the Conservative Party and his wife are expecting a baby in August.

The World Today reports that the leader of the Conservative Party and his wife are expecting a baby in July.

News Update reports that the leader of the Conservative Party and his wife are expecting a baby in March.

DISCONFIRMING EVIDENCE:

The Daily News reports that the leader of the Conservative Party and his wife are expecting a baby in November.

The Globe reports that the leader of the Conservative Party and his wife are expecting a baby in December.

The World Today reports that the leader of the Conservative Party and his wife are expecting a baby in January.

News Update reports that the leader of the Conservative Party and his wife are expecting a baby in March.

Figure 3.6: The three levels of the evidence manipulation for the two arguments used in Experiment 1b.

3.8.2 Results and Discussion

99 participants completed the experiment. Analyses were based on two dependent measures – a rating of argument strength (the truth of the claim), and a rating of the reliability of the sources. As in Experiment 1a, all statistical analysis was conducted on residual ratings, whilst all graphed data is of raw, untreated means.

3.8.2.1 Ratings of Argument Strength

Figure 3.7 displays the mean ratings of argument strength obtained in each experimental condition. An ANOVA was conducted on ratings of argument strength with Evidence (confirms/mixed/disconfirms) and Class (scientific/non-scientific) as independent variables.
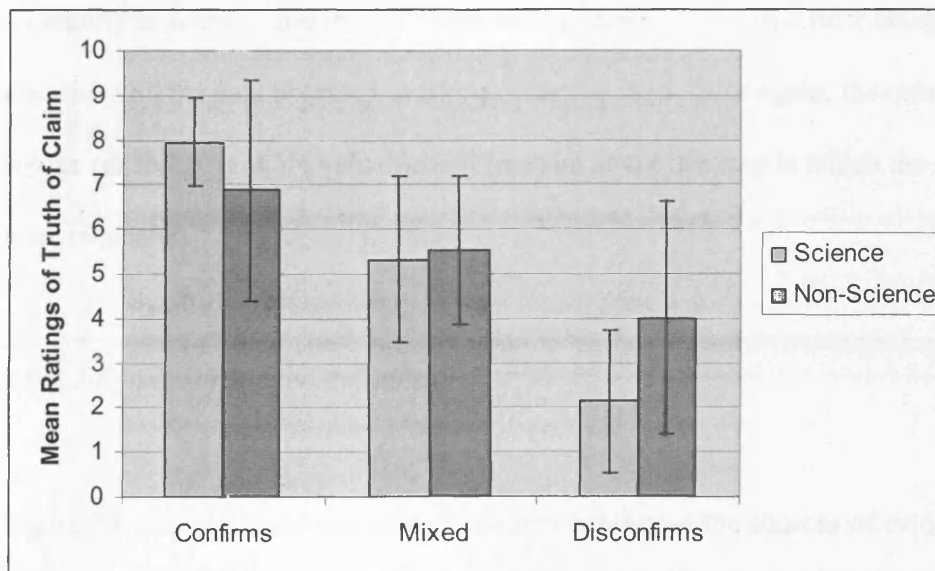

Figure 3.7: Mean ratings of argument strength. Error bars indicate one standard deviation.

There was a main effect of Evidence in the expected direction, $F(2,194) = 79.08$, p

$<.001$. Confirmatory evidence was perceived as stronger ($M = 2.43$, $SD = 1.7$) than

mixed evidence ($M = 0.52$, $SD = 1.71$) which was perceived as stronger than

disconfirming evidence ($M = -1.61$, $SD = 2.05$). Tukey post-hoc tests confirmed that

these differences were significant for each level of the manipulation, and for both the

scientific and non-scientific arguments. There was no main effect of Class, although

there was an interaction between Class and Evidence, $F(2, 194) = 5.68$, $p < .01$.

Scientific argument strength ratings seemed once again to be polarised – more

convincing when based on disconfirming evidence. However, pairwise comparisons

indicated that neither of these differences was statistically significant.

The argument strength ratings support the Bayesian prediction that the more evidence

a claim has, the more compelling it will be. The trend towards polarisation of the

science arguments supports the idea that people are intolerant of uncertainty and

ambiguity in science, and expect 'facts and certainties' in scientific messages, and

also fits with the data obtained in Experiment 1a. And, once again, the ratings of

source reliability provide valuable information about the way in which the arguments

were evaluated.

3.8.2.2 Ratings of Source Reliability

Figure 3.8 displays the mean ratings of the reliability of the sources of evidence in

each condition of the experiment. An ANOVA was conducted on the residuals of

these ratings, with Evidence (confirms/mixed/disconfirms) and Class (scientific/non-

scientific) as independent variables. There was a main effect of Evidence, $F(2,194) =$

23.35, $p <.001$. As predicted, however the ordinal pattern of results is different to the argument strength ratings. While confirmatory evidence produced the highest ratings of source reliability ($M = 1.36$, $SD = 2.12$), ratings of source reliability were higher for disconfirming evidence ($M = -0.14$, $SD = 2.29$) than for mixed evidence ($M = -0.99$, $SD = 2.07$). There was also a main effect of Class in the reliability ratings, $F (1,194) = 87.39$, $p <.001$. Specifically, scientific sources were consistently rated as *more* reliable ($M = 1.86$, $SD = 1.38$) than non-scientific claims ($M = -1.19$, $SD = 2.07$).
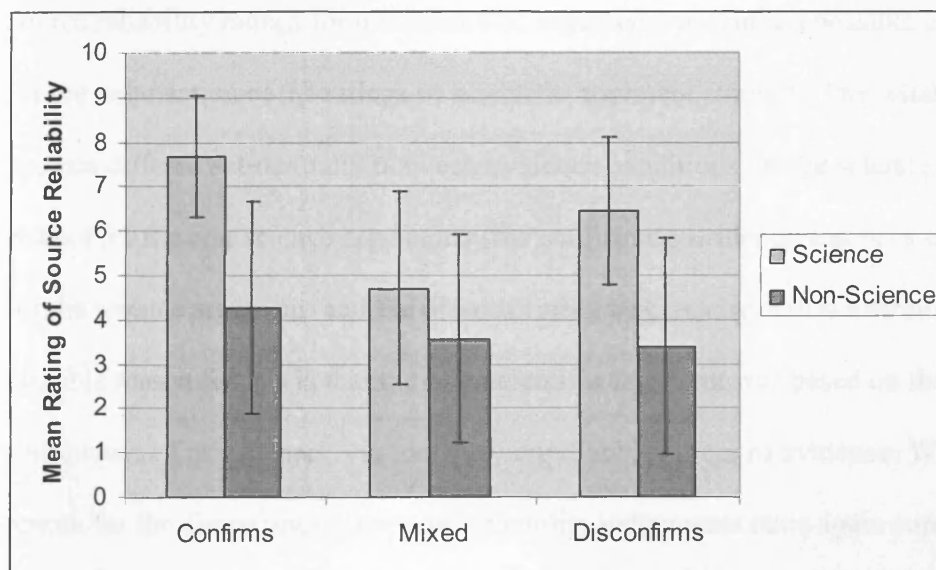
Figure 3.8: Mean ratings of the reliability of the sources of evidence. Error bars indicate one standard deviation.

Tukey post-hoc tests on the Evidence main effect revealed that for the science argument, each level of the evidence manipulation was significantly different, such that ratings of source reliability were *significantly* higher for disconfirming than for mixed evidence. For the non-science argument, only the confirmatory evidence condition differed from the mixed and disconfirming conditions. The mixed and

disconfirming conditions did not differ significantly from each other. The interaction between Evidence and Class was non-significant.

Figure 3.8 shows, however, that the source reliability ratings for the non-scientific arguments were consistently low – even when all the evidence confirmed the hypothesis, ratings of source reliability did not reach the mid-point of the response scale. Another way of describing this pattern of data is to say that the likelihood ratios of the sources in the non-scientific argument were all low, and all fairly similar – even the coherent evidence was not deemed especially diagnostic. This 'squashing' of the source reliability ratings for non-scientific arguments provides a possible explanation for the polarisation of the ratings of scientific argument strength: The reliability of the sources differed substantially between evidence conditions for the science arguments, but not for the non-science arguments (the confirming evidence was *very* confirming for the science argument, and the disconfirming evidence *very* disconfirming). One possible reason for this is that the non-scientific argument was based on the testimonies of newspapers – notoriously unreliable sources of evidence. Whatever the reason for the discrepancy, however, reliability judgements once again supported a Bayesian analysis of argument strength ratings, as well as providing broad support for Bovens and Hartmann's (2003) Bayesian analysis of evidential coherence.

The results of this experiment provide a detailed examination of how people evaluate scientific and non-scientific evidence, under a variety of different conditions. Although I have suggested that the observed differences between scientific and non-scientific arguments can be explained within a Bayesian framework (i.e. scientific and non-scientific argument strength ratings only differ because the ratings of source

reliability do), the analysis has implications for the communication of science. Lack of evidential coherence (i.e. a 'mixed message') reflects badly on judgements of source reliability, and this effect seemed to be strongest for the scientific argument. In both experiments 1a and 1b, a degree of polarisation was observed in argument strength ratings, driven by judgements of source reliability. Combined with the lack of tolerance for scientific uncertainty that seems prevalent among the general public, the results suggest that contradictory scientific evidence may impact badly on perceptions of scientists themselves – and that in a rational, Bayesian fashion, this will feed in to the evaluation of scientific arguments.

3.9   Experiment 1c – Evidence and Belief Change

Thus far, I have examined different aspects of argument and evidence evaluation using the framework of Bayesian theory to guide experimental design. But at root, Bayesian theory is about *belief change* – Bayes' Theorem provides a rational model for incorporating new evidence into your existing beliefs. In the next experiment, I will examine how different types of evidence impact on belief change, in the context of scientific and non-scientific arguments.

The history of science is often portrayed as a series of discoveries – advancements in knowledge and expertise that proceed in an orderly and linear fashion (Gregory & Miller, 1998). Sociologists of science, however, tell a different story. As Collins and his colleagues (e.g. Collins & Pinch, 1993) have repeatedly demonstrated, the progression of science traces a much more haphazard path than this traditional view suggests. When Einstein developed the theory of relativity and introduced his famous

'constant', many new predictions were made and existing theories had to be revised. Typically, Einstein's moment of insight is presented as exactly that – a singular discovery that proved that only the speed of light remains constant across space, time and mass. Authors like Collins and Pinch (1993) document a very different process, however, whereby the scientific orthodoxy is resistant to change and is unwilling to yield to the implications of the new discovery. Counter experiments are run and alternative hypotheses offered – as they should be in a functioning scientific community. The history books, though, record only that a discovery was made, and that our knowledge was updated. This fits with the 'facts and figures' interpretation of science that is taught in schools and presented in the media.

A crucial aspect of the scientific method is the ability to falsify hypotheses (c.f. Popper, 1959), and scientists should always be ready, in the face of replicable and reliable data, to abandon even their most entrenched views about the world. This is one of the factors that distinguishes science from ideology. But if, as Collins and Pinch suggest, even scientists have difficulty accepting a change that has far reaching consequences, are ordinary people really willing to accept scientific evidence if it appears to contradict so completely their existing knowledge?

People are invested in their existing knowledge – a concept known as 'embeddedness' (see, e.g., Quine, 1969). If you were to discover, for example, that the plankton on the bottom of the ocean were a darker shade of brown than you had previously thought, you would be unlikely to need to adapt to this knowledge – most people's knowledge about the colour of plankton is typically not firmly embedded. If, however, you were to discover that the plankton had legs and could walk about, your knowledge about

the taxonomic structure of the world and the place of plankton within it would be shaken. That plankton do not have legs is a fact that is sufficiently embedded in most people's knowledge that changing it would have an enormous knock-on effect on the rest of their knowledge base. Embeddedness is not something unique to scientific knowledge – it can equally be applied to the social domain (Prislin & Oullette, 1996). But I have characterised the public representation of scientific knowledge as a series of facts, and the accepted history of science as a linear advancement of knowledge. How then, do people evaluate scientific, as opposed to non-scientific evidence that does not conform to their expectations?

The next experiment was designed to explore how people update their belief in an initially implausible hypothesis, as they receive increasing amounts of evidence. Bayes' Theorem is, of course, an account of belief revision, and we would expect to see significant changes in participants' assessment of an implausible hypothesis from their prior (which should be low) to their posterior degree of belief. Whether there are any systematic discrepancies in the evidential impact of scientific, as opposed to non-scientific evidence is the focus of the experiment.

A distinction was drawn, however, between science as a *product* (i.e. arguments about scientific topics) and science as a *process* (i.e. arguments based on evidence acquired using a scientific methodology). Two scenarios were developed in which an implausible scientific or non-scientific hypothesis was introduced and evidence in support of that hypothesis incrementally reported to participants. The evidence was obtained either using a scientific method or a non-scientific method. The factorial combination of these two variables (class of argument/method of obtaining evidence)

allowed their relative influence on participants' degree of belief in the implausible hypothesis to be distinguished. By using fictional topics (whether or not a solution released red gas when mixed with a metal/whether the sweets from a sweet shop were edible or not) it was also possible to ensure that the 'embeddedness' of the knowledge in each scenario was equivalent – participants had no reason to hold on to their prior beliefs more in one scenario than in the other, unless, of course, the presence of scientific or non-scientific information encouraged them to do so.

3.9.1   Methods

Participants

99 undergraduate students from the psychology department at Cardiff University took part in the experiment in exchange for course credit.

Design and Materials

Experiment 1c was a between-participants repeated measures design. Two between-participants variables were manipulated (scientific/non-scientific class of argument, scientific/non-scientific method of obtaining evidence), and four measures of argument strength were recorded. First, participants were asked to indicate their initial belief, on a scale of 0 – 10 in a scientific or a non-scientific claim (these were designed such that they would be perceived as implausible by participants, and as such should attract low ratings of initial belief, leaving room for belief in the claim to *change* following the addition of incremental evidence). The scientific claim was:

You have never seen a metal added to a solution cause red smoke before, but you want to know whether it is true for Solution X. How likely do you think it is that adding a metal to Solution X will cause red smoke?

The non-scientific claim was:

You have just seen a sweet shop called the Candy Man. It looks very dirty, and the window displays are yellowed from the sun. You want to know whether the sweets it sells are nice. How nice do you think the sweets will be?

Having indicated their prior degree of belief in one of these claims, participants then read about one, then ten, and finally fifty pieces of evidence that supported the claim. Participants either read that the evidence had been obtained using a scientific method or using a non-scientific method. In the scientific condition participants received the evidence in this format:

You try one sweet and it tastes nice/you add a metal to Solution X and it causes red smoke.

In the non-scientific condition participants received the evidence in this format:

One person tells you that the sweets in the shop are nice to eat/one person tells you that adding metal to Solution X causes red smoke.

After each new piece of evidence, participants were required to indicate their belief in the original claim. Thus four measures of argument strength were obtained (based on prior belief, one, ten and fifty pieces of evidence), as well as an indication of belief change.

3.9.2 Results

Figure 3.9 shows the mean ratings of belief in the claim for each experimental condition and across each level of the evidence manipulation.
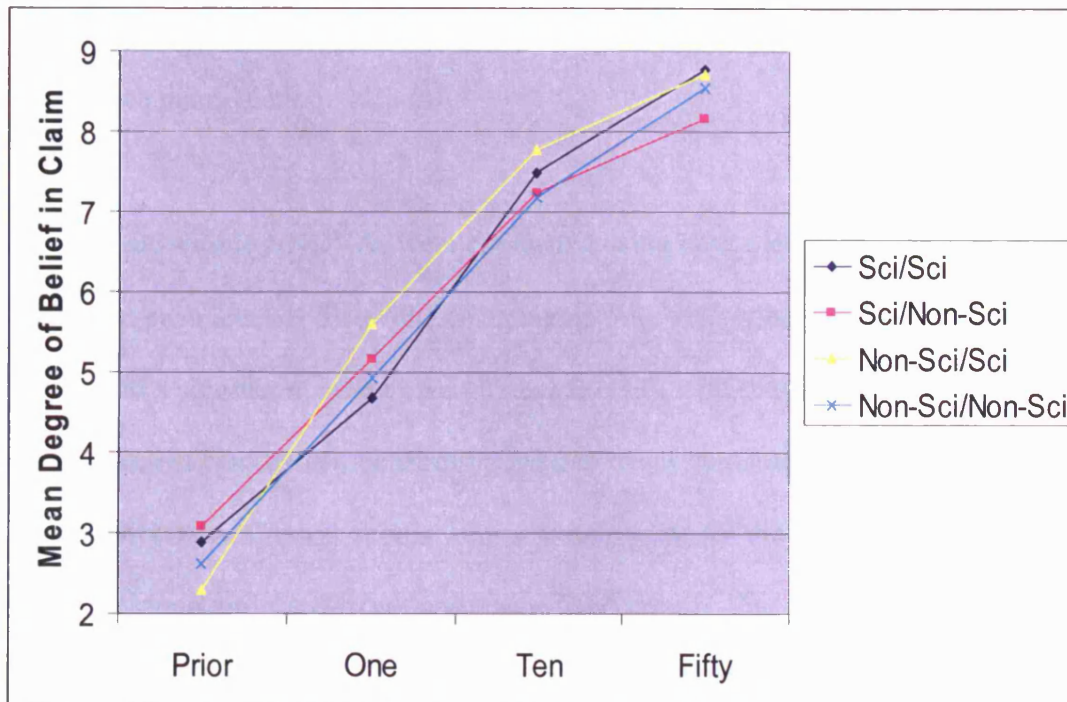
Figure 3.9: Mean ratings of belief in the claim for each experimental condition across each level of the evidence manipulation (Key: Class of Argument/Method of Obtaining Evidence)

A repeated measures ANOVA with Class (scientific/non-scientific) and Method (scientific/non-scientific) as the between-participants variables, and Evidence (prior/one/ten/fifty) as the within-participants variable was conducted on ratings of belief in the claim. As expected, there was a strong main effect of the Evidence manipulation, $F(3, 285) = 514.34$, $p < .001$, such that belief in the claim increased as incremental evidence was received. Prior belief in the claim ($M = 2.71$, $SD = 1.69$)

was lower than belief in the claim following one (M = 5.09, SD = 2.04), ten (M = 7.41, SD = 1.84) and fifty (M = 8.55, SD = 1.48) pieces of evidence. Using this analysis, neither the Class, nor the Method variable had a significant effect on ratings of belief in the claim. There was a significant interaction, however, between the Class and the Evidence manipulation, F (3, 285) = 2.76, p < .05. This can be observed in Figure 3.9: The experimental conditions trace marginally different trajectories across the evidence manipulation.

Four separate simple ANOVAs were conducted using both the between-participants variables at each level of the evidence manipulation. Neither between-participants variable had a significant effect on argument strength ratings at any of the levels of the evidence manipulation – confirming the null result obtained using the repeated measures ANOVA. Change scores were also calculated for the differences between 'prior-one', 'one-ten', 'ten-fifty', and 'prior-fifty' ratings. The only change score where a significant effect of an experimental variable was present was the difference between 'prior-one' ratings, where the non-scientific topic produced significantly more belief change (M = 2.81, SD = 1.94) than the scientific topic (M = 1.94, SD = 1.51), F (1, 95) = 6.29, p < .05.

Overall, belief change occurred in a way consistent with Bayesian predictions – the more evidence for the hypothesis, the more favourably it was evaluated. Some small differences between the scientific and non-scientific conditions did emerge, but it is difficult to draw a strong conclusion from these differences. Certainly, there do not seem to be any *systematic* differences in the way that scientific and non-scientific claims were evaluated. This suggests that belief updating, and the incorporation of

new evidence is broadly similar for scientific and non-scientific arguments – and that educational programmes aimed at improving critical thinking and the understanding of evidence are likely to improve the effectiveness of science communication even if they are not specifically targeted at scientific knowledge. The Bayesian approach predicts that rational evidence evaluation and belief updating should be applied to the evaluation of both scientific and non-scientific arguments, and the results of this experiment provide empirical support for this prediction.

In the final experiment in this chapter, I introduce a variant of the Bayesian approach which will be used extensively in subsequent chapters – Bayesian decision theory – and examine another type of argument commonly found in science communication: Consequentialist arguments.

3.10   Experiment 1d – Consequentialist Arguments

Consequentialist arguments are simply arguments about consequences – that is, arguments about what will happen in the future, *if* a particular action is taken, e.g. 'if you listen to loud music, then you may lose your hearing'. They are, therefore, conditional arguments, as they take the form of 'if P then Q' statements. Consequentialist arguments are very familiar from scientific discourse – climate change is only the most pertinent example of a scientific topic where the consequences of (in)action play a central role in the debate. In fact, in addition to the prevalence of scientific arguments from ignorance, many 'appeals' based on scientific data take the form of a consequentialist argument (e.g. healthy eating or anti-smoking campaigns). There is an extensive psychological literature on how people use and

interpret conditional arguments (see Evans & Over, 2004, or Oaksford & Chater, 2003, for overviews). In particular, research has focused on whether people's use of conditional arguments conforms to a logical or probabilistic normative standard. My focus here, however, is not on people's aptitude for using conditional probabilities, but on what factors might make consequentialist arguments strong or weak in the context of scientific and non-scientific topics.

So far, I have described the Bayesian approach to argument strength as being primarily concerned with subjective probabilities. As Figure 3.2 shows, one's belief in a hypothesis should increase as evidence in support of it grows – a process captured by Bayes' Theorem. And, there is a sense in that consequentialist arguments can be conceived of as simply probabilistic arguments – how *likely* is it that you will lose your hearing, if you listen to loud music?

Consequentialist arguments are not just about probabilities, however – they are also about the desirability of the consequence that they predict. For example, you may consider it highly unlikely that listening to loud music will cause you to lose your hearing, but find the prospect of losing it so terrible that even a small chance of it happening would be unacceptable. A consequentialist argument with a sufficiently undesirable outcome may still be perceived as a strong argument – because the utility of the outcome is so negative.

The way to accommodate this feature of consequentialist arguments within the Bayesian framework is Bayesian decision theory (Edwards, 1961; Keeney & Raiffa, 1976; Savage, 1954), which provides a guide to decision making in situations where

outcomes are uncertain, based on the subjective probabilities, but also the subjective utilities involved (see Chapter 1, and further discussion in Chapter 4).

If you thought, therefore, that there was a high probability of it raining, and it was very important to you that you did not get wet (i.e. there was strong negative utility associated with getting wet), it would be rational for you to take an umbrella. Conversely, if you found it particularly cumbersome to carry an umbrella, had no great aversion to getting wet, and thought it unlikely that it would rain, it would make more sense to leave the umbrella at home. Applying decision theory to consequentialist arguments, the more (subjective) negative utility there is associated with a consequence, the stronger that consequentialist argument should be (for related work applying decision theory to conditional reasoning, see Over, Manktelow & Hadjichristidis, 2004).

The framework of Bayesian decision theory was used to guide the construction of the arguments in the current experiment. The specific focus was on the utility component of decision theory – that is, how bad the consequence posited in each argument was. A consequentialist argument positing a very negative outcome should be perceived as stronger than a similar argument positing a less negative outcome, as illustrated by the following two examples:

"In order to reduce the risk of your car being stolen, you should purchase an alarm system"

"In order to reduce the risk of your car being scratched, you should purchase an alarm system"

Preventing your car being stolen is presumably a stronger reason for installing an alarm system than preventing your car being scratched – and so arguments with more negative outcomes should be stronger. In addition, however, the type of action that is required in order to avoid this negative outcome is also important to the strength of the argument:

"In order to reduce the risk of your car being stolen, you should install an alarm system at the cost of £200"

"In order to reduce the risk of your car being stolen, you should install an alarm system at the cost of £20"

Intuitively, the second argument seems more acceptable than the first. The sacrifice required to avoid the negative outcome in the first argument is much greater, so it is perhaps no surprise that we intuitively find the second argument stronger. It is often the case in real world contexts, however, that the avoidance of negative outcomes requires a degree of personal sacrifice. This variable was therefore also included, in order to enable a more detailed picture of the factors that contribute to the utility component of an evaluation of a consequentialist argument. Are evaluations of consequentialist arguments based on outcome negativity, the amount of sacrifice required to avoid the negative outcome, or an interaction between both these factors? Experiment 1d was designed as an initial attempt to answer these questions.

Scientific arguments often take the form of a consequential or conditional argument – a particularly topical example is climate change. Many scientific arguments about climate change are based on the consequences that our current actions will have for future generations. We may be warned, for example, that if the global climate continues to increase in temperature, glacial ice will melt at an accelerated rate, sea levels will rise and low lying homes will be flooded. This is certainly a negative consequence, but avoiding it might require personal sacrifices that many are unwilling to make. Are scientific consequentialist arguments evaluated any differently to their non-scientific counterparts?

3.10.1 Methods

Design

According to Bonnefon and Hilton (2004), consequentialist arguments may be persuasive (i.e. arguments about how we can obtain a positive outcome), or dissuasive (i.e. arguments about how we can avoid a negative outcome). Following this classification system four dissuasive consequentialist arguments were developed – two relating to scientific topics, and two relating to non-scientific topics.

Three experimental variables (Outcome Utility, Level of Sacrifice and Class) were manipulated at two levels (moderately/very negative outcome utility, small/big sacrifice and scientific/non-scientific argument) across four different argument topics, creating a total of sixteen distinct arguments using a Latin Square Confounded design (see Experiment 1a). All participants were presented with four consequentialist arguments, each concerning a different topic, and were required to provide a rating of

argument strength for each argument on a numerical scale from 0 (Very unconvincing) – 10 (Very convincing). In addition, participants were required to indicate how bad they thought each outcome was, and how bad it would be if they had to make the sacrifice prescribed by each argument, on a scale from 0 (Very bad) – 10 (Not at all bad). Participants saw only one argument from each topic, and participated once in each experimental condition. The topics of the arguments and the order they were presented in were randomised for each participant.

Materials and Procedure

Each participant received an experimental booklet containing four consequentialist arguments on different topics. The two scientific topics were: (S1) some predicted effects of climate change, and (S2) the risks associated with high blood pressure. The two non-scientific topics were: (NS1) the potential benefits of installing a car alarm, and (NS2) the risk of an alarm clock failing to go off in the morning. Each topic is shown below in Figure 3.10. Information pertaining to the experimental manipulation of Outcome Utility and Level of Sacrifice is shown in bold.

S1

The Intergovernmental Panel on Climate Change (IPCC) have claimed that if global warming continues at the current rate, it will cause global sea levels to rise and **thousands of people who live in low-lying areas will lose their homes/tourism will be disrupted**. The IPCC has calculated that if **everyone switched to using energy efficient light bulbs/we all stopped using aeroplanes** the amount of CO2 saved would stop the sea level from rising. To prevent sea levels from rising, and **thousands of people from losing their homes/tourism being disrupt**ed, everyone should **switch the light bulbs in their houses to energy efficient ones/stop using aeroplanes**.

S2

The British Medical Association has issued a statement about the relationship between eating salt and high blood pressure. If people eat over 20g of salt a day, their blood pressure increases and they are more likely to **have a heart attack/have high blood pressure**. Doctors have established that **never eating any processed food/never eating crisps** is an effective way of keeping under the 20g a day limit. To keep under the 20g of salt a day limit, and reduce their chance of **having a heart attack/having high blood pressure**, people should **never eat any processed food/never eat crisps**.

NS1

Imagine you are setting an alarm clock **to wake you up the next morning/to wake you up in time for an important exam**. Your friend warns you that the batteries she has given you for your alarm clock have almost run out of power – she doesn't know how much longer they will last. If the batteries run out of power in the night, you will **sleep late the next morning/sleep late the next morning and miss the exam**. In order to get new batteries, you will have to walk **2 minutes down the road/3 miles into town** to buy some more. Unless you walk **2 minutes down the road/3 miles into town** to buy some new batteries, you will **sleep late/you will sleep late and miss the examination** tomorrow.

NS2

Imagine that you are worried about your car **being stolen/being scratched**. In order to stop people **stealing/scratching** your car, you can install an alarm system that warns people when they approach your vehicle to keep their distance, at a cost of **£20/£200**. In order to reduce the risk of your car **being stolen/being scratched**, you should install an alarm system at the cost of **£20/£200**.

Figure 3.10: The four topics used in Experiment 1d. **Outcome Negativity** and **Level of Sacrifice** information is in bold.

3.10.2 Results and Discussion

All 100 participants completed the consequentialist argument task, providing ratings of the strength of four separate arguments – one from each experimental condition and covering all four topics. The mean ratings of argument strength in each experimental condition are displayed in Figure 3.11 (data from both scientific and non scientific arguments). As in Experiment 1a, statistical analyses were conducted on the residuals of these ratings (Kirk, 1995).
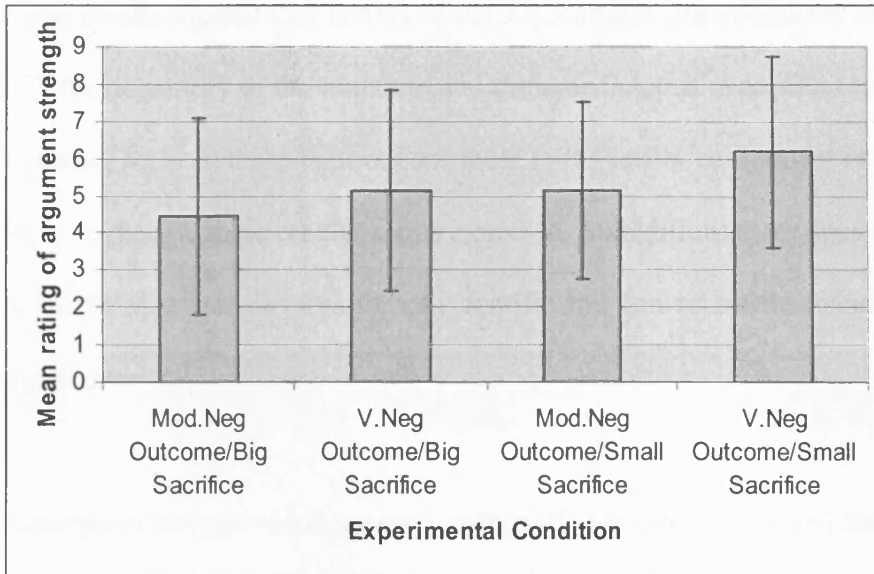
Figure 3.11: Mean ratings of argument strength in each experimental condition (data for both scientific and non-scientific arguments). Error bars indicate one standard deviation.

A three-way analysis of variance (ANOVA), with Outcome Negativity, Level of Sacrifice and Class (scientific of non-scientific) as independent variables was conducted on these residual ratings of argument strength. There was a significant effect of Outcome Negativity, $F(1, 392) = 20.51$, $p < .001$, with more negative outcomes producing higher ratings of argument strength ($M = 0.46$, $SD = 1.96$) than less negative outcomes ($M = -0.45$, $SD = 2.12$). Level of Sacrifice also had a significant effect with smaller sacrifices leading to higher ratings of argument strength ($M = 0.44$, $SD = 2.06$) than bigger sacrifices ($M = -0.43$, $SD = 2.04$), $F(1, 392) = 18.66$, $p < .001$. There was no effect of Class on ratings of argument strength, $p > .05$. In addition, none of the interaction terms between these three variables were significant.

These results suggest that ratings of consequentialist arguments are influenced by both the negativity of the outcome and the sacrifice that is required to avoid this outcome. As both these factors contribute to the utility component of Bayesian decision theory, these results are as expected. In addition, there does not seem to be any difference in the evaluation of scientific and non-scientific consequentialist arguments.

Participants also provided separate ratings of Outcome Utility and Sacrifice Desirability – specifically, participants were asked "how bad would it be if Outcome X occurred?", and "how bad would it be to make Sacrifice Z?". These ratings were also transformed into residuals for the purposes of statistical analysis. A two-way ANOVA with Level of Sacrifice and Outcome Utility as independent variables, and Outcome Negativity residual ratings and Sacrifice Desirability residual ratings as dependent variables was conducted. Only Level of Sacrifice had a significant effect on Sacrifice Desirability ratings, with more negative ratings of the sacrifice required when it was big (M= 4.62, SD= 2.8) than when it was small (M= 7.63, SD= 2.49), F $(1, 392) = 192.04, p < .001$.

Interestingly, however, both Outcome Utility, F $(1, 392) = 220.16, p < .001$, and Level of Sacrifice, F $(1, 392) = 4.67, p < .05$ had a significant effect on Outcome Negativity ratings. Specifically, in addition to the expected effect of the manipulation of outcome negativity on ratings of outcome negativity, participants rated the outcome as significantly *less negative* when the sacrifice required on their behalf was great (M = 0.17, SD = 1.92) than when it was small (M = -0.17, SD = 1.82). This is illustrated in Figure 3.12.
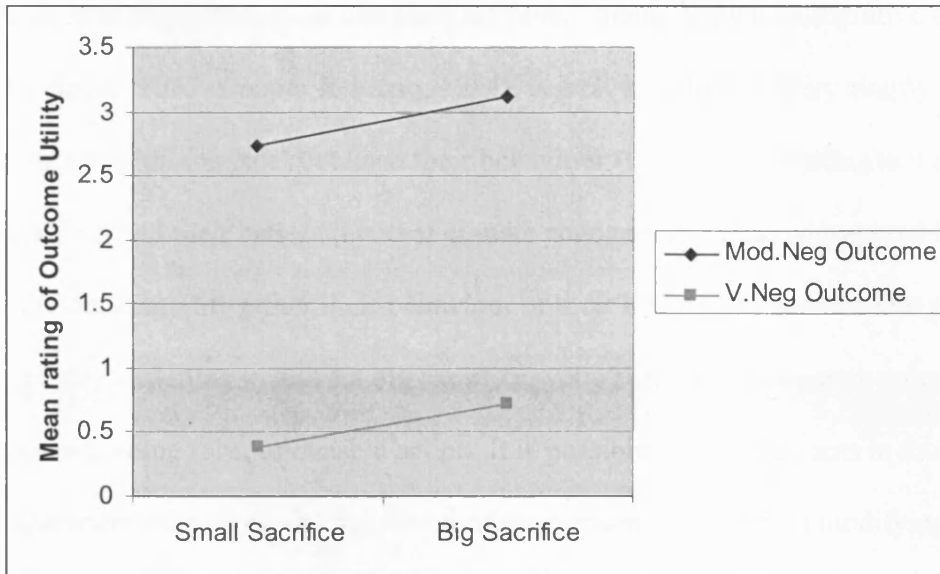
Figure 3.12: The effect of Outcome Utility and Level of Sacrifice on ratings of

Outcome Utility.

Participants seemed to be reducing the negativity of the outcome when avoiding that

outcome required a big sacrifice on their part. On a Bayesian account of decision-

making, the level of sacrifice required would not be expected to feed into the

perceived negativity of the outcome[8]. This unpredicted effect of sacrifice on outcome

negativity was not confined to the evaluation of either scientific or non-scientific

arguments – and so once again, there is no evidence to suggest anything *unique* in the

evaluation of scientific arguments. However, consideration of a potential explanation

for this unpredicted effect has important implications for the communication of

appeals based on scientific evidence.

---

[8] In decision theory, utilities are assumed to be independent, such that changing the context in which a particular outcome is presented should not influence the utility associated with that outcome. Only then can they satisfy the 'impartiality condition', whereby "If you prefer one gamble to another in one state, then you prefer it in all states..." (Allingham, 2002, p44), a requirement for the full subjective utility property. In the current experiment, the utility of a negative outcome should be unaffected by presenting it alongside a small or large sacrifice (for other examples of context effects that violate the independence requirement of decision theory see Stewart, Chater, Stott & Reimers, 2003).

In the social psychological literature the phenomenon known as cognitive dissonance, (Festinger, 1957; Cooper & Fazio, 1984), is well established. Very simply, in order to reduce the 'dissonance' between their behaviour (i.e. being unwilling to make a large sacrifice) and their beliefs (i.e. that climate change may cause widespread flooding), individuals modify either their behaviour or their beliefs. We assume that people are typically unwilling to make a big sacrifice, yet a sufficiently negative outcome demands some form of evasive action. It is possible that participants in this experiment minimised the negativity of the outcome, rather than modifying their behavioural intentions – as a less negative outcome carries a reduced obligation to avoid it.

If consequentialist arguments such as these do engage dissonance mechanisms in this way, then there are important implications for the communication of scientific appeals using the consequentialist argument format. On the one hand, communicators must be cautious in invoking too negative an outcome and linking it to personal sacrifice. Changing environmental behaviour may be more effectively achieved by emphasising how important small sacrifices can be, rather than offering an extremely negative outcome as a reason for making a greater sacrifice. Indeed, research on the use of fear appeals in persuasive communication suggests that there is a danger of inducing defensive reactions if the severity of the message is too high (de Vries, Ruiter & Leegwater, 2002), and that simply increasing severity does not necessarily add to the persuasive impact of a message (Hoog, Stroebe & de Wit, 2005).

On the other hand, however, cognitive dissonance is a well-understood psychological phenomenon – and as such there is a wealth of psychological literature (see, e.g.,

Cooper & Fazio, 1984; Dickerson, Thibodeau, Aronson, & Miller, 1992; Kantola, Syme & Campbell, 1984) that could potentially be brought to bear on the communication of scientific consequentialist appeals. On the current evidence, there is no reason to suggest that there is anything unique about scientific consequentialist arguments. The present results do, however, provide a possible explanation for why communicating science using consequentialist arguments might not be as straightforward as it first appears.

3.11   General Discussion

In this chapter I have aimed to achieve two related goals. Firstly, I have applied the Bayesian approach to informal argumentation to the analysis of socio-scientific arguments – the sorts of representations of science that the general public might evaluate in their daily lives. Secondly, I have reported the results of four experiments that examine different aspects of the way in which people evaluate scientific arguments and evidence.

On the basis of these exploratory experiments there is very little evidence to suggest that people are anything other than Bayesian in their evaluation of scientific arguments and evidence. In all four experiments, participants' responses matched Bayesian predictions to a significant extent. With only one exception (Experiment 1d, see below), the differences that emerged between the evaluation of scientific and non-scientific arguments seemed to be attributable to variation in participants' judgements of factors – such as source reliability – that the Bayesian approach predicts *should* alter ratings of argument strength. That is, although people's subjective estimates of

the parameters that determine argument strength sometimes differed depending on whether the arguments were scientific or not, these differences were *subjectively rational* on the Bayesian account of argument strength. Variation in the perceived strength of scientific arguments is determined by the same factors that determine the strength of non-scientific arguments. From the perspective of designing programmes aimed at improving the communication of science these results are encouraging – there seems to be nothing systematically different about the way in which scientific arguments are processed, at least not on a basic, cognitive level.

In Experiment 1d I reported the only evidence of non-Bayesian reasoning – the perceived negativity of the outcome of consequentialist arguments seemed to be influenced by the sacrifice required to avoid that outcome. In Bayesian decision theory utilities are combined such that a negative outcome and a big sacrifice *should* create a strong consequentialist argument. The utilities associated with the sacrifice and the outcome are assumed, however, to be independent – that is, the magnitude of the sacrifice should not influence the perceived negativity of the outcome (and vice-versa). A negative outcome should be a negative outcome, regardless of the sacrifice required to avoid it (see Allingham, 2002; Stewart, Chater, Stott & Reimer, 2003).

Tellingly, however, this deviation from Bayesian predictions – that is, the only time that participants seemed to be doing anything other than evaluating the arguments rationally – was not specific to scientific arguments. Rather the effect was a general one – although with some important implications for the communication of scientific information that is presented in a consequentialist argument format. Many scientific appeals take the form of a consequentialist argument and the present results suggest

that these appeals may be vulnerable to dissonance effects. However, as dissonance is

a well-established psychological phenomenon there is potentially a wealth of

literature than speaks directly to the issue of effectively communicating scientific

consequentialist appeals (for related discussions of using dissonance theory to

encourage environmentally responsible behaviours, see, e.g., Dickerson, Thibodeau,

Aronson & Miller, 1992; Kantola, Syme & Campbell, 1984).

Taken together, the four experiments reported in this chapter suggest that people

evaluate scientific and non-scientific arguments in a broadly similar way. Of course, a

null effect (i.e. observing no influence of scientific/non-scientific topic on ratings of

argument strength) is not sufficient to establish that *no* difference exists. My claim is

not that there are no differences between scientific and non-scientific argument

evaluation, but that where evaluations of scientific and non-scientific arguments

differed, variation in general factors such as source reliability seemed to explain these

discrepancies. But why should these differences in the reliability of scientific and

non-scientific sources be present in the first place?

A comprehensive answer to this question is beyond the scope of the present research.

I have tentatively suggested, however, that the way in which science is typically

perceived by the general public – in particular, the tendency to expect facts and

certainties, and the corresponding impact this expectation has on judgements of

'uncertain' evidence, or weaker sources – is likely to be important. However, it is

only the application of a content-based, normative framework that permits questions

about the relative strength of scientific and non-scientific arguments to even be posed

in the first place. If the experimental data in this chapter cannot explain why the

general public are intolerant of uncertainty in science, it can at least eliminate the possibility that people evaluate scientific arguments and evidence in a systematically biased fashion.

The main contribution of this chapter lies therefore not in the data obtained, but in the successful application of the Bayesian framework to scientific argument evaluation – something which I believe is novel, and which allows future research to proceed with some precise and systematic questions in hand. By using the tools of Bayesian theory to quantify and measure concepts that are frequently discussed in the literature on scientific arguments, I have been able to offer a means of comparing scientific argument evaluation to non-scientific evaluation that is both systematic and consistent across different experimental tasks. Crucially, while previous studies of informal scientific arguments have focussed on identifying and classifying the *type* of arguments that are used (see, e.g., von Aufschaiter et al 2008), the Bayesian approach allows analysis at the level of argument *content*. And because the Bayesian approach is *normative* I have been able to make some assessment of how *good* people were at evaluating scientific arguments (in comparison to Bayesian norms for argumentation, and their evaluation of non-scientific arguments). The questions posed and the initial data obtained were contingent on applying a framework that permitted key variables in the analysis of informal scientific arguments to be quantified and measured.

While the Bayesian approach may not be sufficient to provide a complete account of how science is evaluated, it does at least provide a method for posing systematic questions. By adopting this framework, I have been able to formulate questions that would simply not be possible without. An understanding of informal argument

strength seems a prerequisite for understanding science communication, and in this sense, the Bayesian approach has substantial heuristic value. Whether or not the Bayesian approach turns out to be the *best* way of analysing informal scientific arguments, it is minimally somewhere to start.

3.12 Chapter Summary

The purpose of this chapter was to investigate the broad question of how people evaluate scientific and non-scientific arguments. Motivated by the urgency with which some socio-scientific messages must be communicated in the public domain, I sought to apply a Bayesian framework in order to explore people's judgements of scientific and non-scientific argument strength, and their evaluations of scientific and non-scientific arguments. A fundamental problem facing researchers interested in the evaluation of scientific arguments is the lack of a content-based, normative framework with which to study scientific argument evaluation. By introducing the framework of Bayesian theory and using it as a guide for the design of these experiments, I was able to formulate some precise questions about how people evaluate scientific arguments, quantify and measure some important variables that impact on the convincingness of scientific arguments, and draw some conclusions about how the data obtained might relate to the broader issues of science communication and the public evaluation of science.

The results of these experiments raise many questions that cannot be answered here, and that would require a broader methodological approach to tackle comprehensively. It is also important that these data should be replicated with other populations that

might reasonably be expected to differ in their evaluations of science (practising

scientists, for example, or science communicators – see Chapter 6 for further

discussion). However, as an exploratory attempt to identify the factors that dictate

people's evaluations of scientific arguments and evidence, this chapter provides some

important markers to guide future empirical questions – and a theoretical framework

within which to pose them.

Having used the Bayesian framework to pose systematic questions about a

contemporary topic in argumentation (how people evaluate socio-scientific

arguments), the second part of my experimental research focuses on a longstanding

puzzle in the philosophical literature: Slippery slope arguments (SSAs). This fits with

my stated goal of using the Bayesian approach to examine phenomena in

argumentation that have not necessarily received an empirical treatment.

Chapter 4: A Bayesian analysis of Slippery Slope Arguments

"...Slippery slopes are metaphors. While metaphors can be helpful, they often start by enriching our vision and end by clouding it...One can always shout 'slippery slope!' but without more details this is hardly an argument at all." (Volokh & Newman, 2003, p23).

4.1   Chapter Overview

Having introduced the Bayesian approach to informal argumentation, defended it as a normative theory of argument strength, and used it as a framework for studying people's evaluation of scientific arguments and evidence, I will now apply the Bayesian approach to a long-standing puzzle in philosophy: The slippery slope argument (SSA). SSAs have a bad philosophical reputation. They seem, however, to be widely used and frequently accepted in many legal, political, or ethical contexts. Theories of persuasion have for a long time recognised that a message does not necessarily have to contain a strong argument in order for it to be effective. But are SSAs simply fallacies of reason – arguments that are 'wrong' but persuasive? Using rational criteria for distinguishing strong and weak SSAs proposed in Hahn and Oaksford (2006a, 2007a), I provide experimental evidence that people's evaluations of SSAs match normative predictions (Experiments 2-5). In the course of examining SSAs, many issues relating to the social psychological literature on persuasion are raised. I have attempted to relate the findings in this chapter to existing social psychological research wherever possible and in the final experiment (Experiment 6),

I show that SSAs that are stronger according to normative, Bayesian, criteria are also more persuasive and give rise to greater attitudinal changes.

4.2 Introduction

The 'slippery slope' is an intuitive metaphor that is used to refer to a class of arguments with a distinctive form, but varied content. Classified as a fallacy of reason by most critical thinking textbooks (Woods, Irvine & Walton, 2004) and philosophers (e.g., Enoch, 2001), yet frequently used and widely accepted in applied domains such as politics (van Der Burg, 1991), law (Lode, 1999) and bioethics (Lamb, 1988; Launis, 2002), the slippery slope argument is a controversial topic in the field of argumentation. For most, the argument possesses the somewhat undignified status of "wrong but persuasive", and therefore fits neatly into the category of arguments that argumentation theorists call fallacies (although see Corner & Hahn, 2007, and Hahn & Oaksford, 2007a). Of course, the notion that a message may be persuasive without necessarily containing a 'strong' argument is not a notion that is confined to philosophical argument analyses.

Decades of research in the social psychological literature on persuasion and attitude change has established that a systematic analysis of message content is not necessarily required in order for a message to be persuasively effective. In fact, leading theories of persuasion make an explicit distinction between 'message' and 'source' factors (e.g., Petty & Cacioppo, 1984), such that a persuasive appeal may be entirely determined by, for example, the perceived expertise of a message giver, rather than an analysis of the message itself. The concept of a fallacious argument is one that is built

on the normative, philosophical notion of 'acceptable' argument types. Persuasiveness is something that dictates how effective an argument actually is – and in practice, SSAs seem to enjoy a prevalence and persuasiveness that belies their philosophical reputation.

It would seem, then, that there is a potential conflict between the philosophical standing of SSAs and their use in practical contexts. Consideration of the following examples suggests, however, that SSAs are not all as persuasive, or as fallacious, as each other:

(1) "If we allow gay marriage, then in the future people will want to marry their pets".

(2) "If voluntary euthanasia is legalised, then in the future there will be more cases of 'medical murder'".

(3) "If we accept voluntary ID cards in the UK, we will end up with compulsory ID cards in the future".

These examples exhibit an enormous amount of variation in their plausibility and persuasiveness. Few would agree that homosexual unions are the beginning of the slippery slope to inter-species marriages, although precisely this argument has been put forward by a group called the American Family Research Council (2001). Perhaps, then, SSAs deserve their bad philosophical reputation. The euthanasia example seems more plausible, although not sufficiently plausible to prevent the Dutch Government from legalising certain forms of voluntary euthanasia. Finally, it

seems extremely likely that ID cards in the UK will become compulsory once they have been introduced – in fact, if they are to function as an effective security measure, this may be a necessity. From the dubious logic of (1), through the calculated risk-taking of (2), to the almost inevitable consequence of (3), SSAs display an impressive variation in their convincingness. It may therefore not be useful to treat SSAs as a generic concept that can simply be labelled a "fallacy", as some seem very compelling. In other words, from the fact that there are weak SSAs, it doesn't necessarily follow that SSAs are fallacious. By the same token, however, it does not seem desirable to licence all SSAs as acceptable arguments, and the mere fact that an argument is persuasive seems insufficient grounds for claiming that it is 'strong'.

In this chapter I outline rational criteria (first proposed in Hahn & Oaksford, 2006a, 2007a) for distinguishing SSAs. By applying insights from the psychological literatures on persuasion, argumentation and human judgement, I suggest that although SSAs may be strong or (in the case of the American Family Research Council) very weak, there is nothing inherently *wrong* with them. By analysing SSAs from a Bayesian perspective it is possible to identify the factors that determine their strength. I present the results of five experiments that suggest that (like any other form of argument) there is a great deal of variation in how convincing people find individual instantiations of the argument to be. More importantly, this variation can be experimentally controlled and theoretically predicted.

Whilst it is simple enough to produce an intuitive characterisation of SSAs, they have resisted attempts to provide a comprehensive definition. As Rizzo and Whitman (2003) put it, "there is no paradigm case of *the* slippery slope argument" (Rizzo &

Whitman, 2003, p 544). Authors have typically opted either to differentiate multiple independent forms of SSA (e.g. Walton, 1992b) or to treat only a very select group of arguments as genuine examples of SSAs (e.g. Govier, 1982). Walton (1992b), for example, distinguishes four types of SSA, suggesting that some SSAs involve causal mechanisms ('causal' SSAs), some set precedents ('precedent' SSAs), while others are attributable to the vagueness of concepts and categories ('sorites' SSAs). A fourth type combines features from each of these SSAs ('full' SSAs). Other authors have opted to avoid such a detailed taxonomy, choosing instead to list 'core features' that SSAs generally seem to possess. Rizzo and Whitman (2003), for example, identify three components they claim are common to all SSAs; (1) an initial, seemingly acceptable, decision; (2) a 'danger case' that is clearly unacceptable; and (3) a process or mechanism by which the initial decision will raise the likelihood of the danger case.

I will attempt, however, to give a definition of SSAs that is useful from a psychological perspective – that is, useful specifically for a psychological analysis of SSA strength. SSAs are a particular breed of *consequentialist* argument (see Experiment 1d, Chapter 1, of this thesis, and also Hahn & Oaksford, 2007a; for analysis of SSAs as consequentialist arguments see Walton, 1992b; Oakley & Cocking, 2005; for recent experimental work on other forms of consequential conditional see Bonnefon & Hilton, 2004; Thompson, Evans & Handley, 2005). A dissuasive consequentialist argument (or deterrent) warns against a particular course of action on the grounds that it will lead to an undesirable outcome, or consequence. An SSA, however, posits not only a negative outcome but the idea that this outcome might in the future be *re-evaluated as positive*, if an initial proposal goes ahead.

A general consequentialist argument, might oppose the legalisation of cannabis because it would lead to an increase in smoking related respiratory problems. A slippery slope argument would oppose legalisation on the grounds that attitudes towards harder drugs might become more positive in the process, and in the future a substance like cocaine might also become legal. This gives the slippery slope four distinct components:

1. An initial proposal (A)

2. An undesirable outcome (B)

3. The belief that allowing (A) will lead to a re-evaluation of (B) in the future

4. The rejection of (A) based on this belief[9]

The alleged danger lurking on the slippery slope is the fear that a presently unacceptable proposal (B) will (by any number of psychological processes – see, e.g., Volokh, 2003) in the future be re-evaluated as acceptable. If we withhold the right to free speech from a neo-Nazi organisation, what will prevent us from censoring legitimate political dissent in the future? The proponent of this argument is inherently appealing to the malleability of public opinion to reject an otherwise appealing course of action. The uncertainty of the future is such that any reasoning about it is at best presumptive. Yet SSAs trade on the uncertainty of the future, and appear to be acceptable in a number of contexts (Lode, 1999; Volokh, 2003). In light of the fact that there has been no empirical investigation of the slippery slope, a pressing task is to examine if, when, and how SSAs can be strong arguments.

---

[9] This SSA definition is not designed to correspond to any one of the types of SSA identified by Walton (1992b), although it might best be thought of as embodying the 'precedent' and 'sorites' SSAs in Walton's taxonomy. This point is discussed further in Chapter 6.

The re-evaluative nature of the SSA links the argument to a classic persuasive phenomenon – the 'foot in the door' (FITD) technique, first identified by Freedman and Fraser (1966). As the name suggests, the FITD technique is a persuasive method that involves making a small request at first, based on the assumption that once an individual has agreed to this small request (i.e. opening the door), they will be more likely to subsequently accept a more substantial request. It is this substantial request that is the actual goal of the persuasion attempt – the small request is simply a means of obtaining this end. One of the explanations proposed for the FITD effect is a process of re-evaluation, such that the substantial request is re-evaluated as more acceptable once the small request has been agreed to. There are clear parallels between FITD and our characterisation of the SSA, and there is already existing work outlining the effectiveness of FITD appeals. Interestingly, however, a meta-analysis of FITD studies found that FITD appeals are *only* effective when they relate to pro-social topics (Dillard, Hunter & Burgoon, 1984) – that is, they are only effective when the ultimate goal of the persuasive communication is to obtain a positive outcome. The implication of this finding is important, because SSAs are typically about *avoiding* a negative outcome, rather than procuring or obtaining a positive outcome. An empirical study of the SSA is distinct, then, from the FITD literature, although clearly importantly related.

4.3   Capturing SSA strength

I have cast the SSA as a particular type of dissuasive consequentialist argument. What is required, then, is a framework with which to *evaluate* SSAs – a method of distinguishing strong SSAs from weak ones. As discussed in Chapter 1, the Bayesian

approach to the analysis of informal arguments extends probabilistic approaches to

scientific reasoning and rationality (see e.g., Howson & Urbach, 1996; Oaksford &

Chater, 2001) to everyday argumentation (Hahn & Oaksford, 2007a; Hahn, Oaksford

& Corner, 2005; Oaksford & Hahn, 2004; see also Korb, 2004). This approach seeks

to interpret reasoning patterns as probabilistic changes in subjective degrees of belief,

and applies probability theory as a normative framework for evaluating consistency

and change in degree of belief. Having applied this framework to the analysis of

scientific arguments in Chapter 3, I will now conduct an experimental investigation of

the SSA using the Bayesian approach.

Using a Bayesian model of argument strength an SSA should be viewed as strong

(and therefore convincing) to the extent that its consequences seem *probable* given

the available evidence. The idea that the putative outcomes of SSAs are not as

probable as their proponents claim is central to its reputation as a scare-mongering,

fallacious argument. As Oakley and Cocking (2005) observe;

"It is widely recognised that the standard problem with many slippery slope

arguments is that they fail to provide us with the necessary compelling evidence that

significantly worse circumstances will actually come about..." (Oakley & Cocking,

2005, p232).

Clearly, an important component of SSA strength will be the likelihood of the

undesirable outcome it predicts *actually occurring*. In one sense, therefore, an SSA

can be analysed as the simple conditional probability $P(B|A)$ – that is, what is the

chance of (B) occurring given (A)? Consequently we should expect SSAs whereby

the initial proposal is likely to bring about the feared outcome to be stronger than ones where that probability is low. An account of SSA strength would be incomplete, however, if the *utilities,* or values, associated with its components were ignored. In particular, philosophers interested in applied domains such as law or bio-ethics where SSAs are popular have implicitly recognised that probabilistic *and* utilitarian concerns are crucial determinants of consequential and slippery slope argument acceptability (e.g., Holtug, 1993; Lode, 1999). This distinguishes SSAs from most other fallacies of argumentation (for overviews of the traditional catalogue of fallacies see e.g., Woods et al. 2004).

Perhaps the most important aspect of SSAs is that they advocate *decisions*, and as such are not just arguments about factual claims. In Chapter 1, I noted that Bayesian decision theory (see, e.g. Edwards, 1961; Keeney & Raiffa, 1976; Savage, 1954) is a normative framework for decision-making in situations where outcomes are uncertain, based on the subjective probabilities and utilities involved. And, in Chapter 3, I reported the results of an experiment using Bayesian decision theory to examine how people evaluate scientific and non-scientific consequentialist arguments.

According to this theory, agents should seek to maximize subjective expected utility (i.e., potential gain) in their choices. This means that different agents can rationally choose different courses of action if their respective assessments of probabilities and utilities differ. However, there is still a normative standard in operation here, in that the evaluation of decisions by a given rational agent must be derivable from more fundamental valuations – namely the probabilities and utilities they assign. Specifically, the subjective expected utility of a decision (SEU) must correspond to

the probability-weighted sum of the utilities associated with a particular course of action. Equation 4.1, shown earlier in Chapter 1, is repeated here, and shows the SEU for any number of outcomes (x_i), where (P) is the subjective probability and (U) the subjective utility of each outcome:

$$\sum_i P(x_i)U(x_i) \qquad \text{(Eq. 4.1)}$$

Given then, that SSAs advocate particular actions based on their putative consequences, this framework can be brought to bear directly on the assessment of their strength. The higher the probability that some feared outcome will be bought about by an initial course of action and the greater the utility of avoiding this feared outcome, the stronger the SSA will be. This link between SSA evaluation and an assessment of expected utility has been noted informally by philosophers in the past. Holtug (1993), for example, claimed that in relation to SSA strength, "the more probable the causal connection is, and the more we want to avoid (B), the stronger the argument" (Holtug, 1993, p404). The tools of decision theory provide a formal framework for these intuitions.

The idea of combining probabilistic and utilitarian information in order to derive an account of message strength links with other research within social psychology. Fishbein and Ayzen's (1975) theory of reasoned action states that an attitude (A) is a function of the sum of an individual's beliefs (b_i) about an attitudinal object's attributes multiplied by the expected values (e_i) of these beliefs (i.e. a summed evaluation of the attributes of the attitudinal object):

135

$$A = \sum b_i e_i \quad \text{(Eq. 4.2)}$$

In assessing a sportswoman, for instance, we might evaluate her on a number of dimensions – her endurance, her speed, and her skill level perhaps. The degree to which we believe she possesses these attributes is said to define our attitude towards her – a sum of expected values. Our attitude towards her is therefore based on a set of evaluative beliefs that when altered, should have a corresponding effect on our attitude.

Mathematically, this model is a generalization of decision theory (in the sense that it can incorporate Equation 4.2 as a special, limiting case). As a result of this generalization, however, the normative status associated with decision theory is lost[10]. The theory of reasoned action makes no claim to provide a normative account of how people's attitudes *should* change if their attitudes are to qualify as rational – it is a descriptive theory of how they *actually* do. Moreover, similarly to the probabilogical models (discussed in Chapter 1), its theoretical focus is on how the manipulation of one attitude will propagate through the rest of the attitudinal system. It is not an account of argument strength.

To clarify these distinctions further, 'attitude' is a psychological construct that refers to an implicit tendency or disposition which is expressed through evaluation of a particular entity with some degree of favour or disfavour (Eagly & Chaiken, 1993).

---

[10] The Bayesian framework treats probabilities as subjective degrees of belief, which fits naturally with expectancy value theory. However, probabilities must also obey certain constraints specified by the axioms of probability theory in order to be rational. The same is true of utilities (see e.g., Savage, 1954). No such constraints are imposed in the theory of reasoned action.

Arguments, by contrast, are one way of altering attitudes. They are only one way of many, because attitudes can be formed in several ways – for example, through direct experience of the attitudinal object. Moreover, as outlined above, persuasive messages can function through a variety of non-content based factors other than the arguments they might contain. Nevertheless there are many persuasive contexts in which argument strength is the central factor in eliciting persuasion (e.g., Eagly & Chaiken, 1993; Petty & Caccioppo, 1984; Kruglanski, Fishbach, Erb, Pierro & Mannetti, 2004), yet there is to date no developed theory of what makes an argument weak or strong. This is a widely acknowledged gap within the literature on persuasion (Areni & Lutz, 1988; O'Keefe & Jackson, 1995; Petty & Wegener, 1991; van Enschot-Van Dijk, Hustinx & Hoeken, 2003). It has been circumvented in practice by using experimenter intuition to derive 'weak' and 'strong' sample arguments and confirm their respective status through pre-testing. Moreover, the bulk of the literature on persuasion has used the *same* topic and associated sample arguments –concerning the introduction of 'comprehensive exams' for students (c.f. Petty & Cacioppo, 1986) – as experimental materials. Hence the development of a *theory* of argument strength should be of central concern to researchers studying persuasion.

I seek here to argue for the Bayesian decision-theoretic framework as providing such a theory in the realm of consequentialist arguments. That it links with existing social psychological work on attitude change is encouraging. It is important, however, to keep separate the concepts of argument and attitude, and I will return to their relationship later in the chapter.

With these theoretical considerations in hand, the first experiment in this chapter

sought to show in detail how the decision theoretic framework provides an account of

SSA strength. Specifically, it sought to test the extent to which its characterization of

the relative strength of different SSAs matched participants' subjective evaluations.

4.4 Experiment 2

Participants were required to read several short scenarios containing slippery slope

arguments, and provide a rating of argument strength (as illustrated in Figure 4.1).

The experiment was designed to demonstrate experimentally that slippery slope

arguments vary in convincingness, and that this variation can be predicted by

manipulating (i) the probability, and (ii) the utility of the predicted future outcome, in

line with the predictions of Bayesian decision theory.

Regarding (i), an argument where the probability of the outcome (B) given the initial

proposal (A) is high should be more convincing than an argument where $P(B|A)$ is

low. In the present experiment, the conditional probabilities presented to participants

were varied by describing either a probable or an improbable mechanism by which

the proposed outcome of the argument could occur.

Regarding (ii), a predicted outcome is a necessary component of slippery slope

argumentation, but predicted outcomes that are perceived to have only a moderately

negative expected utility will not be "feared" or avoided as much as outcomes with

very negative expected utility. Predicted outcomes with very negative utilities will

provide a stronger argument against the proposed course of action. In the present

experiment, the outcome utilities of the arguments presented to participants were set as either moderately negative or very negative.

Figure 4.1 shows an example scenario as seen by participants in each condition of the experiment. In the first version of the scenario the probability of the outcome (B) given the initial proposal (A) is high (because of the alleged difficulty of formulating clear medical guidelines), whilst the utility of the predicted outcome is very negative (in the form of an increase in involuntary euthanasia). In the second version the probability of this negative outcome occurring is designed to be lower. In the third version the predicted outcome is less negative (other patients on the ward feeling less comfortable knowing that euthanasia is taking place), but probable. Finally, in the fourth version the outcome is both less negative and less probable.

(Likely/Very Negative Outcome)

Whilst flicking through a copy of Ethics magazine, you come across an article on the thorny issue of euthanasia. Despite almost unanimous agreement (from both the medical profession and terminally ill individuals) on the acceptability of helping some patients to end their suffering, opponents claim that the legalisation of voluntary euthanasia will lead to an increase in cases of involuntary euthanasia – or "medical murder". The British Medical Association has warned that once voluntary euthanasia is permitted in some cases, it will be difficult to formulate clear guidelines about when doctors can euthanize patients. The article ends with the view of the author about the future of euthanasia legislation;

"We should oppose the legalisation of euthanasia in the UK, as it will lead to an increase in the number of instances of 'medical murder'".

(Unlikely/Very Negative Outcome)

Whilst flicking through a copy of Ethics magazine, you come across an article on the thorny issue of euthanasia. Despite almost unanimous agreement (from both the medical profession and terminally ill individuals) on the acceptability of helping some patients to end their suffering, opponents claim that

the legalisation of voluntary euthanasia will lead to an increase in cases of involuntary euthanasia – or "medical murder". The British Medical Association has indicated, however, that there will be extremely clear and strict guidelines about if and when doctors may euthanize patients, and those who break them will be removed from the medical register. The article ends with the view of the author about the future of euthanasia legislation;

"We should oppose the legalisation of euthanasia in the UK, as it will lead to an increase in the number of instances of 'medical murder'".

(Likely/Less Negative Outcome)
Whilst flicking through a copy of Ethics magazine, you come across an article on the thorny issue of euthanasia. Despite almost unanimous agreement (from both the medical profession and terminally ill individuals) on the acceptability of helping some patients to end their suffering, opponents claim that the legalisation of voluntary euthanasia will lead to other hospital patients feeling that their lives are not as worthwhile. The British Medical Association has warned that once voluntary euthanasia is permitted, terminally ill patients may start to view their lives as of less worth than healthy individuals. The article ends with the view of the author about the future of euthanasia legislation;

"We should oppose the legalisation of euthanasia in the UK, as it will lead to other terminally ill patients feeling psychologically damaged by the process"
(Unlikely/Less Negative Outcome)

Whilst flicking through a copy of Ethics magazine, you come across an article on the thorny issue of euthanasia. Despite almost unanimous agreement (from both the medical profession and terminally ill individuals) on the acceptability of helping some patients to end their suffering, opponents claim that the legalisation of voluntary euthanasia will lead to other hospital patients feeling that their lives are not as worthwhile. The British Medical Association has indicated, however, that most hospital patients are unconcerned by the thought of voluntary euthanasia in hospitals. The article ends with the view of the author about the future of euthanasia legislation;

"We should oppose the legalisation of euthanasia in the UK, as it will lead to other terminally ill patients feeling psychologically damaged by the process"

Figure 4.1: An example scenario from Experiment 2, topic (i). The four versions of the scenario illustrate the four conditions of the experiment.

4.4.1   Method

Participants

60 undergraduate psychology students from Cardiff University participated in

Experiment 2 for course credit.

Design

The experiment was a 2 (probable/improbable mechanism) X 2 (moderately/very

negative outcome utility) factorial design. Both variables were manipulated across

four different argument topics, creating a total of sixteen distinct arguments. All

participants were presented with four slippery slope arguments, each concerning a

different topic, and were required to provide a rating of argument strength for each

argument on a scale from 0 (unconvincing) – 10 (very convincing). The topics of the

arguments and the order they were presented in were randomised for each participant

using a Latin Square Confounded design (see Kirk, 1995; and Experiment 1a in this

thesis).

Materials and Procedure

Each participant received an experimental booklet containing four slippery slope

arguments on different topics. The topics were (i) the legalisation of voluntary

euthanasia, (ii) the distribution of newspapers to a small General Store, (iii) the

introduction of I.D. cards, and (iv) the cessation of postal deliveries to houses

inhabited by vicious dogs. Figure 4.1 shows the arguments used in topic (i).

4.4.2 Results and Discussion

The mean ratings of argument strength from each condition of the experiment are displayed in Figure 4.2.
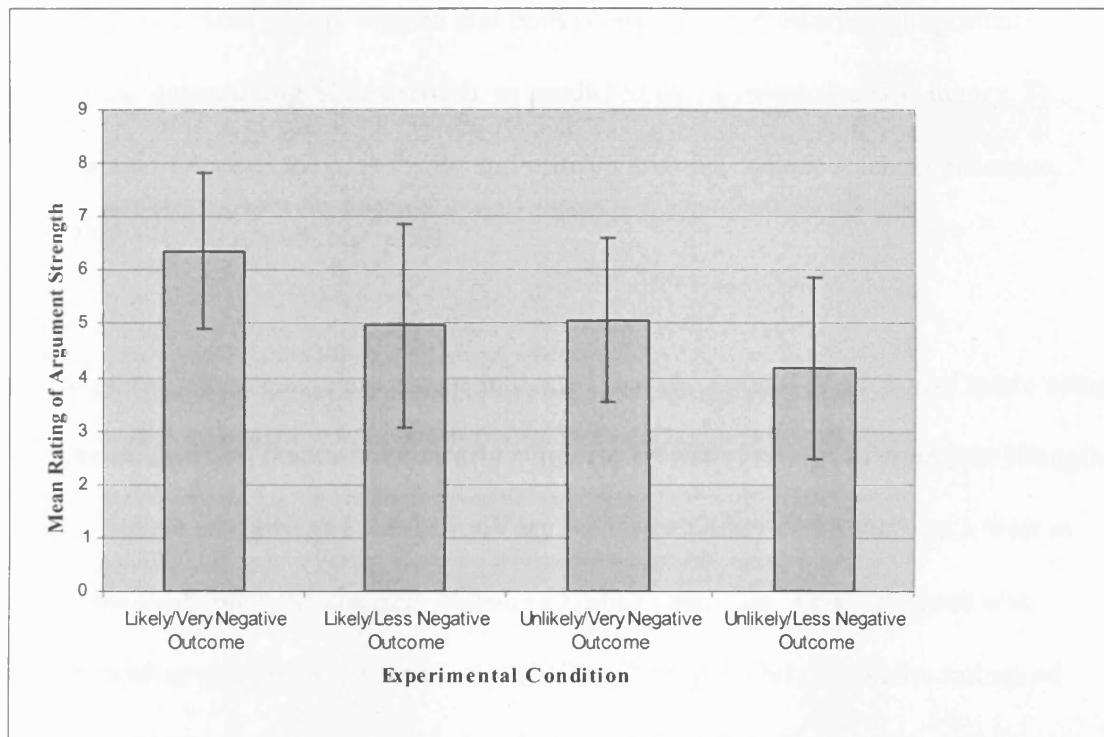
Figure 4.2: Mean ratings of argument strength from Experiment 2. Errors bars indicate one standard deviation.

To statistically analyse data from Latin Square Confounded designs, participant effects within the ratings are factored out and the analyses are conducted on the residuals (Kirk, 1995; see also Experiment 1a of this thesis). Though this changes the absolute numerical values, it leaves the overall shape displayed in Figure 4.2 unaltered. An Analysis of Variance (ANOVA) revealed that the probability manipulation had a significant effect on ratings of argument strength. SSAs with more probable outcomes were rated as significantly more convincing ($M = 0.51$, $SD = 1.82$)

than SSAs with less probable outcomes (M = -0.5, SD = 1.67), F (1,236) = 22.26, p <

.001, eta^2 = .09. In addition, SSAs with more negative outcomes produced

significantly higher ratings of argument strength (M = 0.58, SD = 1.62) than SSAs

with less negative outcomes (M = -0.57, SD = 1.82) , F (1,236) = 28.71, p < .001, eta^2

= .11. These main effects suggest that both probability and utility are important

factors in determining SSA strength, as predicted by Bayesian decision theory. The

interaction between the probability and utility variables did not reach significance, F

(1, 236) = 1.67, p >.05, eta^2 = .01.

Two planned pairwise comparisons indicated that the ordinal predictions I made using

Bayesian decision theory were clearly supported. Firstly, ratings of argument strength

were highest in the High Probability/Very Negative Utility condition, and lowest in

the Low Probability/Moderately Negative Utility condition. This difference was

confirmed as significant with a t-test, t (119) = 7.48, p < .001. Secondly, ratings of

argument strength in the two mixed conditions, where the effects of the variables were

expected to work in opposition, were not significantly different from each other.

In decision theory, probabilities and utilities are combined multiplicatively. This

would suggest that when both factors take on extreme values, a non-linear effect on

ratings of argument strength should be observed – and reflected in a significant

interaction between the two independent variables. Specifically, combining high-

probabilities and extreme utilities should lead to a larger than additive effect on

argument strength. There is a trend to this effect in the data (see Figure 4.2). That the

interaction did not reach significance could indicate one of two things – either

participants were not combining probabilistic and utilitarian information in the

manner prescribed by decision theory, or the experiment did not possess sufficient power for the interaction effect to be observed. The power of the interaction term reported above would favour the second explanation (eta^2 values are reported for ANOVA statistics throughout the experiments in this chapter).

Three replications of Experiment 2 were therefore conducted using exactly the same experimental design and a similar number of participants each time, but varying some of the topics in order to maximise the strength of the manipulations. Introducing new topics also permits a test of the generality of the account.

4.5   Experiment 3

4.5.1   Methods

76 undergraduate psychology students from Cardiff University participated in the experiment for course credit. The design, materials and procedure were identical to Experiment 2, except that an argument about genetic modification replaced topic (iv). The argument either claimed that publishing techniques to genetically modify crops would lead to unchecked attempts to modify the DNA of animals and humans (very negative utility), or that people would become less tolerant of misshapen or discoloured fruit and vegetables, because it would be possible to create perfect products, and would therefore waste more food (moderately negative utility). The outcome probability variable was manipulated by having a panel of experts state that this outcome would either be likely, or unlikely to occur.

4.5.2 Results

An ANOVA was conducted with Outcome Probability and Outcome Utility as the independent variables, and residual ratings of argument strength as the dependent variable. There was a main effect of both independent variables. Arguments with likely outcomes were rated as stronger ($M = 5.8$; $SD = 2.2$) than arguments with unlikely outcomes ($M = 4.4$; $SD = 2.6$), $F(1, 300) = 38.52$, $p < .001$, $eta^2 = .01$. Very negative outcomes produced higher ratings of argument strength ($M = 5.8$; $SD = 2.3$) than moderately negative outcomes ($M = 4.5$; $SD = 2.3$), $F(1, 300) = 34.1$, $p < .001$, $eta^2 = .01$. The interaction term did not reach significance, $F(1, 300) = 1.51$, $p > .05$, $eta^2 = .005$.

4.6   Experiment 4

4.6.1   Methods

80 undergraduate psychology students from Cardiff University participated in the experiment in exchange for course credit. The design, materials and procedure were identical to Experiment 3.

4.6.2   Results

An ANOVA was conducted with Outcome Probability and Outcome Utility as the independent variables, and residual ratings of argument strength as the dependent variable. There was a main effect of both independent variables. Arguments that contained a likely outcome were rated as more convincing ($M = 5.05$; $SD = 2.5$) than arguments that contained an unlikely outcome ($M = 4.3$; $SD = 2.2$), $F(1, 316) = 12.4$,

$p <.001$, $eta^2 = .04$. Arguments containing very negative outcomes were more

compelling ($M = 5.1$; $SD = 2.3$) than arguments containing moderately negative

outcomes ($M = 4.2$; $SD = 2.3$), $F(1, 316) = 11.39$, $p <.001$, $eta^2 = .04$. The interaction

term did not, however, reach significance, $F(1, 316) = .013$, $p >.05$, $eta^2 = .001$.

4.7   Experiment 5

4.7.1   Methods

97 sixth form students from four colleges across South East Wales took part in the

experiment, as part of a National Science and Engineering Week (NSEW) project.

The design, materials and procedure were identical to Experiment 2, except that an

argument about developments in cloning technology replaced topic (ii). This change

was made in order to ensure that the topics of the arguments were more 'scientific'

(fitting in with the NSEW project). The argument either stated that once animals were

cloned, then humans would be cloned in the future (very negative outcome), or that

animals would be cloned to feed people, and that there were unknown dangers with

this approach (moderately negative outcome). The cloning technology was either

stated to be easy to replicate (likely outcome), or difficult to replicate (unlikely

outcome).

4.7.2   Results

An ANOVA was conducted with Outcome Probability and Outcome Utility as the

independent variables, and residual ratings of argument strength as the dependent

variable. There was a main effect of both independent variables. Arguments with

likely outcomes were rated as stronger ($M = 6.05$; $SD = 2.05$) than arguments with

unlikely outcomes ($M = 4.9$; $SD = 2.1$), $F(1, 384) = 35.98$, $p < .001$, $eta^2 = .09$. Very

negative outcomes produced higher ratings of argument strength ($M = 5.9$; $SD = 2.1$)

than moderately negative outcomes ($M = 5.1$; $SD = 2.2$), $F(1, 384) = 15.43$, $p < .001$,

$eta^2 = .04$. This time, however, the interaction term did reach significance, $F(1, 384)$

$= 4.84$, $p < .05$, $eta^2 = .01$. Pairwise comparisons revealed that when the outcome was

very negative, probable arguments were rated as significantly more convincing ($M =$

0.68, $SD = 1.63$) than less probable arguments ($M = 0.02$, $SD = 1.67$), $t(191) = 2.87$,

$p < .01$. Probable arguments were also significantly more convincing ($M = 0.38$, $SD =$

1.59) than less probable arguments ($M = -1.1$, $SD = 2.01$) when outcome utility was

moderately negative, $t(193) = 5.57$, $p < .001$. Very negative/less probable arguments

attracted higher ratings ($M = 0.02$, $SD = 1.67$) than moderately negative/less probable

arguments ($M = -1.1$, $SD = 2$), $t(193) = 4.09$, $p < .001$ but the effect of outcome utility

was not significant for arguments with probable outcomes. This suggests that a strong

probability manipulation, perhaps linked to the combination of topics used in

Experiment 5, may have been driving the significance of the interaction term.

However, as only one topic differed from the original set of materials used in

Experiment 2, it is also possible that the sample (sixth-form students, rather than

undergraduates) influenced the strength of the probability/utility interaction.

4.8   Meta-Analysis

The results of experiments 2-5 were almost identical – a strong main effect of both

independent variables and a weaker interaction term in each case (although in

Experiment 5 the interaction term did reach statistical significance). It is possible that

the reason for the weakness of the interaction terms in experiments 2-5 was an

insufficient number of participants. Certainly, in each experiment, the power associated with the interaction term was small. However, the experiments also differed in terms of the exact topics used in the argument materials. Therefore it is also possible that the significant interaction between probability and utility in Experiment 5 was due to the combination of topics used in the materials for this experiment. In addition, the sample in Experiment 5 were sixth-form students, rather than undergraduates. This may also have had an effect on the interaction term.

Following the guidelines for meta-analytic procedures outlined in Rosenthal (1984), the data from experiments 2-5 were combined (the procedure is based on quantitatively establishing that experiments testing similar hypothesis share enough commonalities in design and experimental power to be legitimately combined). This created a data set containing 313 participants (each providing four estimates of argument strength). A test of heterogeneity (ANOVA) was then conducted in order to establish a) whether the interaction term would reach significance with a larger sample and b) whether the effect of the probability and utility manipulations differed depending on the combination of topics used in each experiment.

Probability and utility were entered as independent variables, as well as a 'dummy' independent variable denoting which of the four experiments the data was obtained in. Residual ratings of argument strength were used as the dependent variable. Strong main effects of both probability, $F(1, 1236) = 101.11$, $p < .001$, and utility, $F(1, 1236) = 85.37$, $p < .001$, were obtained. However, the interaction between probability and utility did not reach significance, suggesting that a lack of experimental power may not be the explanation for the non-significant effect of the probability/utility

interaction. In addition, there was no main effect of the dummy-coded experiment variable. This suggests that the effects of the probability and utility manipulations in each of the experiments were of comparable magnitude.

Despite a substantial increase in the number of participants included in the meta-analysis, the multiplicative effect of probability and utility did not have a significant effect on ratings of argument strength. In Experiment 5, however, the interaction between probability and utility was significant. There are two possible explanations for this result. Firstly, the materials used in experiments 2-5 differed slightly, such that the particular combination of topics used in Experiment 5 may have produced particularly powerful manipulations of probability and utility. Future experiments could establish whether the topics used in Experiment 5 reliably produce an interaction between the probability and utility manipulation, and why. A second, explanation, however, is that the sample in Experiment 5 were sixth-form students, rather than undergraduates. The experiment was conducted as part of a National Science and Engineering Week project designed to engage non-academic audiences in scientific projects. It is possible that the sixth-form sample were more motivated and paid greater attention to the materials, having never participated in a psychological experiment before. This greater attentiveness may have meant that the experimental manipulations were more powerful.

4.9   Experiment 6

Experiments 2-5 established that people's judgments of SSA strength seem to broadly track Bayesian predictions of when these arguments *should* be perceived as strong or

weak. I have sought to distinguish earlier in the chapter the theoretical notions of argument strength and message persuasiveness or effectiveness (the traditional dependent measure in social psychological studies of persuasion and attitude change). Consequently, no attempt was made in these experiments to measure the effect of the arguments on people's attitudes about the topics the arguments related to. However, having offered an account of SSA strength an obvious next question is whether people's perceptions of SSA strength also predict their responses to the arguments as persuasive messages. People may judge SSAs with extremely negative and highly probable outcomes to be strong but nevertheless display little or no attitudinal change in response to them. Whilst this would not necessarily undermine the appropriateness of Bayesian decision theory as a *normative* framework for analyzing SSA strength, my claim that Bayesian theory provides a framework for understanding people's evaluations of SSAs would certainly be strengthened by an indication that a strong SSA (in Bayesian terms) is also a persuasive argument. The next experiment was designed to examine the impact of slippery slope arguments using a different dependent variable – attitude change.

In the persuasion literature, attitude change is the measure most often used to operationalize message effectiveness. There is reason to believe that the SSAs I have developed so far *will* bring about differing amounts of attitudinal change. This is because they explicitly manipulate the probability and utility of the outcome introduced in the argument. Some models of persuasion already assume that attitudes are changed in line with Fishbein and Ajzen's (1975) theory of reasoned action (see e.g., Albarracin and Kumkale, 2003; Albarracin & Wyer, 2001). In the context of these theories, outcome probabilities and utilities are assumed to determine attitudes

under conditions of so-called elaborative processing, which is adopted by highly involved participants (typically because the message concerns them directly). These message recipients are assumed to generate thoughts, prompted by the argument, that ultimately give rise to attitudinal change. It thus seems likely that such change should occur for SSAs, where the outcome and the manipulation of probability and utility is already part of the message itself.

In fact, post hoc analysis reveals that some of the weak and strong arguments used in persuasion research make reference to outcomes that differ in both probability and utility (Areni & Lutz, 1988; Petty & Wegener, 1991; van Enschot-Van Dijk, Hustinx & Hoeken, 2003). This means that despite being concerned with persuasion rather than argument strength, some previous research has tested the issues at stake here (although not using SSAs). The findings, though, have been mixed; while some studies have reported results that are consistent with effects of both probability and utility (Albarracin & Wyer, 2001; Albarracin & Kumkale, 2003), other studies have found effects of utility only (Areni and Lutz, 1988; Johnson, Smith-McLallen, Killeya & Levin, 2004), or have concluded that probabilistic information in persuasive messages is almost impossible to detect for untutored participants (van Enschot-Van Dijk, Hustinx & Hoeken, 2003). Moreover, these results are limited to a narrow range of materials (typically Petty & Cacioppo's 1986 arguments about the introduction of comprehensive examinations). Consequently, a direct test with SSAs themselves is necessary.

One argument topic from the bank used in Experiment 2 was selected for Experiment 6 – the introduction of voluntary I.D. cards. Participants were asked to indicate their

favorability towards the introduction of voluntary I.D. before and after reading one of

two SSAs against the introduction of voluntary I.D. cards. Only the very strong SSA

(where there was a high probability of a very negative outcome occurring) and the

very weak SSA (where there was a low probability of a less negative outcome

occurring) was used. This allowed any difference in attitudinal change to be observed

clearly. It was expected that participants would show significantly more negative

attitudinal change if they had read the very strong SSA than if they had read the very

weak SSA.

4.9.1   Method

Participants

80 undergraduate students from Cardiff University participated in the experiment as

part of a set of experimental tasks, in exchange for a small payment.

Design, Materials and Procedure

Experiment 6 was designed to assess whether SSAs that had been rated as strong and

weak in Experiment 2 produced differential degrees of attitude change. Participants

were required to indicate their initial favorability towards the introduction of I.D.

cards in the UK on an eleven point scale from 0 (extremely bad) – 10 (extremely

good). They then completed several other unrelated tasks (e.g. making similarity

judgments of geometric stimuli), as the experiment was conducted as part of a bigger

package of tasks, lasting approximately 30 minutes. The experimenters were careful

to ensure that the other experimental tasks were indeed unrelated – that is, there was

no other task requiring the consideration of information about I.D. cards. Participants

next read one of two SSAs relating to the introduction of I.D. cards – either a very

strong SSA (with a high probability of a very negative outcome obtaining) or a very

weak SSA (with a low probability of a less negative outcome obtaining) – see Figure

4.3. Once they had read this argument, they were required to indicate again their

favorability towards the introduction of voluntary I.D. cards, on the same 11-point

scale.

You attend a discussion session on I.D. cards and civil rights. You notice that while most people are in favour of the I.D. Card Bill itself, there is strong opposition to it nonetheless. The main problem appears to be that further legislation has been suggested (the Compulsory I.D. Bill) in which the cards are compulsory and it is an offence not to carry your I.D. The offence would be punishable by a £1000 pound fine. The government's legal advisor has indicated that further legislation would be difficult to oppose once the I.D. Card Bill had been passed, as the House of Lords support the idea of compulsory I.D.
(Likely/Very Negative Outcome)

You attend a discussion session on I.D. cards and civil rights. You notice that while most people are in favour of the I.D. Card Bill itself, there is strong opposition to it nonetheless. The main problem appears to be that further legislation has been suggested (the Compulsory I.D. Bill) which makes the cards compulsory to carry at all times. If requested to produce I.D. by a Police Officer, you must do so within 28 days at your local police station. The government's legal advisor has indicated, however, that any further legislation would meet with strong opposition from the House of Lords, who do not support the idea of compulsory I.D.
(Less Likely/Less Negative Outcome)

Figure 4.3: The strong and weak arguments used in Experiment 6.

4.9.2 Results and Discussion

Pre and post-message attitude scores were entered as within-subjects variables into a repeated measures ANOVA, with experimental condition (weak vs. strong message) as the between-subjects factor. There was a significant main effect of pre-post message attitude change, with post-message attitudes significantly less favourable (M = 5.05, SD = 2.51) than pre-message attitudes (M = 5.8, SD = 2.58), F (1, 78) = 14.56, $p <. 001$. There was also a significant interaction between attitude change and experimental condition, F (1, 78) = 6.47, $p <. 05$. Pairwise comparisons revealed that pre-message attitudes in the weak and strong conditions were not significantly different from each other. Post-message attitudes were less favourable following the strong SSA (M = 4.62, SD = 2.47) than following the weak SSA (M = 5.8, SD = 2.58), but this difference did not reach statistical significance either. Additional analyses revealed, however, that the negative change in favorability towards the introduction of voluntary I.D. cards was greater following the strong argument (M = -1.25, SD = 2.09) than following the weak argument (M = -0.25, SD = 1.33). This difference was significant, t (78) = 2.54, $p <.05$.

These data suggest that SSAs that are perceived as strong (using a measure of argument strength) also cause more attitude change. While the Bayesian account as a normative theory of argument strength does not require that greater attitudinal change be associated with arguments of greater strength (there may be any number of reasons why, in practice, a 'strong' argument fails to alter attitudes), these results provide additional support for the notion that a Bayesian account of SSA strength does, in fact, also have some descriptive value. The results of experiments 2-5 suggested that people's judgments of argument strength map broadly on to a Bayesian account of

argument strength. The data in this experiment suggest that if an argument is strong using Bayesian criteria, it is also likely to be persuasive, as measured by attitudinal change. To this extent, people seem inclined to 'follow their own advice' – and take strong arguments to be persuasive messages.

To the best of my knowledge, this is the first experimental demonstration that a Bayesian measure of argument strength is supported by a persuasive measure of attitude change, and therefore represents an important bridge between the typically independent literatures on argumentation and argument strength on the one hand, and persuasion and attitude change on the other. With regards to the latter, these results are interesting because some research has suggested a role for outcome utility only (Areni & Lutz, 1988; Johnson et al, 2004; van Enschot-Van Dijk, Hustinx & Hoeken, 2003), so that finding a clear effect of probability with novel materials and ostensibly 'disinterested' participants is important. Moreover, even the outcome *utility* effects differ from past research in that of the outcomes of the SSAs are unlikely to be of any direct personal consequence to participants (something which is typically held to be important for message content to be processed fully).

4.10   General Discussion

The results of experiments 2-5 are the first empirical demonstration that SSAs vary predictably in their acceptability, and that this variation is broadly captured by a Bayesian account of argument strength. With regard to argumentation theory and the study of the fallacies in general, this is of interest because variation in strength for arguments of identical structure has typically been problematic for existing theories of

fallacy (e.g., van Eemeren & Grootendorst, 2004), but the idea that argument strength is a graded concept is a central tenet of the Bayesian account. In this respect the results mirror those recently obtained for other supposed fallacies such as the 'argument from ignorance' (Hahn & Oaksford, 2007a; Oaksford & Hahn, 2004; Chapter 3 of this thesis).

With regard to the philosophical debate about slippery slope arguments specifically, the results suggest that the credibility that (some) slippery slope arguments possess in applied domains such as law or medical ethics can be justified. The clear implication of the data obtained in experiments 2-5 is that SSAs are viewed as differing in strength, with some arguments seeming far more convincing than others. That they are not simply 'persuasive but wrong' follows from the fact that the key variables involved in their evaluation – probability and utility – have a normative basis in Bayesian decision theory. Further support for this assertion was obtained in Experiment 6, by demonstrating that degrees of strength in SSAs (as defined by the criteria of Bayesian decision theory) seem to correspond to degrees of persuasiveness for SSAs (as defined by a measure of attitude change). This demonstrates that SSAs are not persuasive independently of their content (i.e., it is not the *form* of the slippery slope that is fallacious or otherwise).

The data are also in line with a wide range of theories of persuasion and attitude change (e.g., Eagly & Chaiken, 1993; Kruglanski et al, 2004; Petty & Cacioppo, 1984), which predict that a strong message should cause more attitude change than a weak one. That the normative, Bayesian notion of argument strength and the descriptive, persuasive notion of attitude change seem to be so closely related in SSA

evaluation is encouraging from the point of view of providing a comprehensive and practical, rather than purely philosophical account of SSAs. In fact, while previous work examining the potential contrast between the normative justification and persuasive success of arguments has suggested that they are certainly not mutually exclusive (O'Keefe, 2003, 2005), only very basic and rudimentary dialectical norms (such as making a message clear and understandable) have been shown to correlate with persuasive efficacy. The experiments reported in this thesis therefore add to the growing body of evidence that people's intuitive judgements of argument strength are accurately predicted by Bayesian norms for argumentation (Hahn, Oaksford & Corner, 2005; Hahn & Oaksford, 2006a, 2006b, 2007a, 2007b; Oaksford & Hahn, 2004).

The questions of rationality and normativity addressed in this thesis are familiar to philosophers and cognitive psychologists, and they are central to the literature on argumentation. Argumentation theorists in both psychology and philosophy are concerned with the *evaluation* of arguments, and standards for arguments addressed at reasonable, rational critics. This normative emphasis is apparent in the developmental literature on argumentation (Kuhn & Udell, 2003), in pragma-dialectical theories of argumentation that seek to clarify the procedural norms that govern rational debate (van Eemeren & Grootendorst, 2004), and recent work outlining a formal, Bayesian approach to measuring argument strength (Hahn & Oaksford, 2006a; 2007a).

By contrast, social psychologists researching attitudes and persuasion have been concerned with (seemingly) entirely different, descriptive questions regarding the *actual processes* by which belief and attitude change comes about. This typically does

157

not involve any attempt to evaluate these mental processes as 'good' or 'bad', and often what constitutes persuasive success seems entirely odds with standards of normative soundness in argumentation (although see O'Keefe, 2003, 2005). But are the worlds of argumentation and persuasion really so far apart? Closer inspection reveals that normative questions about argument quality arise even within this descriptive enterprise of trying to characterize how and why messages are persuasive.

This is because 'argument strength' is a factor in *all* process models of persuasion. Competing dual process models such as the Elaboration Liklihood Model (Petty & Cacioppo, 1984) and the Heuristic-Systematic Model (Chaiken, 1980) or the Uni-model of persuasion (Kruglanski et al, 2004) differ in the exact way, and the circumstances under which, the content of a persuasive message and peripheral, non-content factors (such as source and receiver characteristics) combine into an overall persuasive outcome. *All* models, however, assume that message content, and hence 'argument strength' has a role in persuasion. Moreover, for message recipients that are competent and motivated (perhaps because of personal involvement with the issues at hand) the strength of the arguments presented is assumed to be the main factor influencing persuasion. A full characterization of when, why and how persuasive messages effect attitude change is consequently not possible without a theory of argument strength.

Crucially, such a theory requires a standard against which arguments can be compared. Characterizing an argument as 'good' or 'bad', 'strong' or 'weak' is an *evaluative* statement that requires an evaluative standard or norm. This is further apparent in the intuitive characterizations of 'strong arguments' by social

psychologists: Strong arguments are frequently described as arguments that are "logically sound", "defensible" and "compelling" (e.g., Petty & Cacioppo, 1979, p1920). Johnson et al. (2004) speak of 'strong' arguments as "cogent, rational arguments" (p216). That normative terms such as 'rational' and 'logical' should arise in this context is no coincidence. It is a direct consequence of the fact that measuring argument strength requires an evaluative (i.e. normative) standard.

What counts as a 'good argument' *cannot* be settled in a purely descriptive manner. While it is, to some extent, possible to operationalize for practical purposes the distinction between 'strong' and 'weak' arguments as the distinction between arguments that are persuasive and arguments that are not (as has been prevalent in social psychological research – see O'Keefe & Jackson, 1995, for ways in which argument quality is conceptualised in theories of persuasion), this will never suffice to derive a *theory* of argument strength. If all there is to argument 'strength' is 'persuasive effect', then the supposedly empirical statement that 'stronger arguments' give rise to greater persuasion (at least under certain conditions), becomes a tautology that is true by definition.

This is not to criticise the strategy of *deferring* questions of argument strength in persuasion through the use of pre-tested materials that operationally equate strength with effectiveness (see e.g., Petty & Cacioppo, 1986, for discussion). This strategy has been tremendously successful in building our understanding of the processes underlying persuasion. However, it has left social psychologists without an account of one of the main (if not *the* main) determinants of persuasion – argument strength.

Fishbein and Ajzen (1981) stated that the "general neglect of the information contained in the message is probably the most serious problem in communication and persuasion research" (Fishbein & Ajzen, 1981, p359). Similarly, Petty and Cacioppo (1986) noted that one of the least researched and least understood issues in psychology is that of message quality; there had been literally thousands of studies on extra-message factors in persuasion, but virtually no investigation of messages themselves. Yet this situation persists more or less unchanged to this day (Johnson, Maio, & Smith-McLallen, 2005; van Enschot-Van Dijk, Hustinx & Hoeken, 2003). Arguably, this is no accident, but a consequence of the fact that an explicit consideration of normative standards has become increasingly alien to social psychologists over the last few decades.

That argument strength cannot be understood without a normative standard or referent against which arguments can be evaluated has been noted before (Areni, 2002; Areni & Lutz, 1988; O'Keefe, 2003, 2005; O'Keefe & Jackson, 1995). O'Keefe and Jackson, for example, conclude that all present treatments of this variable within the literature suffer from a common underlying flaw, namely "the lack of an independently-motivated normative account of argument strength" (O'Keefe & Jackson, 1995, p91). They suggest, however, that such an account will not be obtained using the traditional normative tool for argument evaluation in philosophy – logic – but will come from dialectical and pragma-dialectical approaches to argumentation (e.g., van Eemeren, Grootendorst, 2004) and related work in informal logic (e.g., Walton, 1989; Woods & Walton, 1982). These approaches focus not on inherent message features, but on underlying "message production principles" that reflect *procedural* obligations in argumentation. Dialectical approaches to argumentation

160

have sought to explicate the procedural norms that govern rational discourse, such as

an obligation to provide support for one's position when challenged. So-called

'fallacies' such as the SSA have typically been viewed as failures to meet one's

procedural 'burden of proof' in rational discourse (see Hahn & Oaksford, 2007b, for

references).

The view that logic is insufficient for a theory of argument strength is supported by

experiments 2-5: Classical Logic simply has nothing to say about the evaluation of

SSAs, because SSAs are not deductive inferences. However, the present studies also

indicate why a pragma-dialectical approach to argument quality will not suffice.

Argument strength or quality is an intrinsic property of arguments that can be

evaluated even in the absence of any surrounding discourse. This is illustrated by a

consideration of the two sample SSAs from the Introduction section of this chapter,

which seem differentially strong, even without further dialectical knowledge of who

generated them, how they were generated, and why:

(1) "If we allow gay marriage, then in the future people will want to marry their pets".

(2) "If voluntary euthanasia is legalised, then in the future there will be more cases of

'medical murder'".

Such content-dependent variation in argument strength is beyond dialectical

approaches to argument strength (see also Hahn & Oaksford, 2007a). Moreover, as I

argued in Chapter 2, establishing violations of procedural obligations such as failing

to meet one's burden of proof depends itself on assessment of argument strength: A

particular claim can fail to meet one's burden of proof only *because* it is weak. Consequently, that procedural violation cannot in turn be invoked to explain the argument's weakness (see also Hahn & Oaksford, 2007b).

There is one other link between the present work and the persuasion literature; though both probabilities and utilities have frequently been discussed in the context of persuasive messages, utilities (frequently under the header of 'valence') have typically been seen as something *external* to the argument or message under consideration, and hence something distinct from argument strength (e.g., Albarracin & Wyer, 2001; Johnson et al, 2004; for a notable exception see Petty & Wegener, 1991) – even in contexts where the arguments themselves have made reference to utilities (Areni & Lutz, 1988; van Enschot-Van Dijk, Hustinx & Hoeken, 2003). The theoretical analyses and experimental results presented so far should demonstrate why for consequentialist arguments such as SSAs this is inappropriate, and both utilities and probabilities are essential ingredients of argument strength. This point is important for persuasion research because the most widely used weak and strong arguments in persuasion studies are those from Petty and Cacioppo's (1986) materials concerning student "comprehensive exams" which are largely consequentialist in nature. Although post hoc analysis of these particular materials has suggested that they do, in fact, differ along dimensions of likelihood and outcome desirability, the extent to which the probabilistic information is detectable by ordinary participants is debatable (van Enschot-Van Dijk, Hustinx & Hoeken, 2003). A clear demonstration that probabilities and utilities combine to predict ratings of SSAs, a particular type of consequentialist argument, is therefore a novel empirical result.

4.11   Chapter Summary

This chapter provides the first empirical evidence pertaining to a question of longstanding philosophical debate: Can slippery slope arguments ever be considered to be acceptable arguments and if yes, why? Bayesian decision theory provides a normative framework for reasoning about behaviour that can be applied to the behavioural choices targeted by consequentialist arguments such as SSAs. This application indicates the conditions under which SSAs are subjectively rational. At a minimum therefore, the experiments in this chapter would seem to provide a concrete rebuttal of the position that using or taking heed of SSAs is *never* a good idea. This position has been articulated recently by Enoch (2001) who claimed that even though it is possible to construct SSAs that are strong and compelling, people are inherently poor at abiding by the distinction between good and bad SSAs. Thus there is a 'meta' slippery slope between 'good' SSAs and 'bad' SSAs, and the use of 'good' SSAs will trigger a process that will ultimately lead to the spread of 'bad' SSAs. According to Enoch (2001), this means that even good SSAs are ultimately bad. Experiments 2-6 suggest, however, that people have no difficulty in consistently distinguishing between strong and weak SSAs – and that the distinctions they make are subjectively rational according to the normative model of Bayesian decision theory (see also Corner & Hahn, 2007). A significant contribution of the work in this chapter is therefore in providing an empirical answer to the question of whether it can *ever* be rational to be persuaded by a slippery slope argument.

Argumentation theorists interested in developing a normative metric for argument strength (e.g., van Eemeren & Grootendorst, 2004), cognitive psychologists studying

consequential and conditional reasoning (e.g., Evans & Over, 2004; Oaksford & Chater, 1998), and persuasion researchers concerned with developing the notion of argument strength beyond its current position (e.g., Areni & Lutz, 1988; Petty & Wegener, 1991; van Enschot-Van Dijk, Hustinx & Hoeken, 2003) typically have little to say to each other, despite covering extremely close conceptual ground (for notable exceptions, see O'Keefe, 1995, 2003, 2005). This analysis of the SSA would seem to provide an ideal focal point for a much needed interdisciplinary integration.

In the next chapter, I extend my analysis of SSAs by considering a possible mechanism on which they may be based: Similarity-based categorisation. I will aim to show that not only is it subjectively rational to be persuaded by SSAs (in terms of Bayesian argument strength) – subjectively strong SSAs may be predicated on a simple, *objective* measure of similarity. Specifically, when the beginning and the end of a slippery slope are more similar, and are therefore perceived as belonging to the same category, the SSA will be stronger.

Chapter 5 – Similarity-based Categorisation: A mechanism for SSA Evaluation

5.1   Chapter Overview

In Chapter 4, I reported the results of five experiments demonstrating that people evaluate SSAs in line with the subjectively rational criteria of Bayesian decision theory. In addition, arguments that were strong from a Bayesian perspective were also persuasive from the point of view of attitude change. This suggests that SSAs can be subjectively rational, and that subjectively rational SSAs are also persuasively effective. However, the ultimate rejection of the "wrong but persuasive" tag that plagues SSAs would be provided by a demonstration that the differential convincingness of SSAs has some *objective*, empirical basis. Therefore in addition to identifying and manipulating the factors that dictate subjective SSA strength, it is important to ask whether people have good reason to be persuaded by at least some slippery slope arguments. In other words, are there reasons to believe that 'slippage' actually occurs in the real world? In this chapter, I describe the results of three experiments aimed at elucidating a mechanism on which evaluations of SSA may be based: Similarity-based categorisation.

5.2   Mechanisms of the SSA

It is often claimed by those authors that have been positive about SSAs that conceptual vagueness (e.g. the difficulty of providing a precise definition of "terminally ill") and a fear of constructing arbitrary distinctions (e.g. deciding which

terminally ill patients' lives are "worthwhile") provides the rationale for many SSAs

(e.g., Lode, 1999, p1499). For example, Govier (1982) suggests that the process of

psychological assimilation acts as a catalyst for slippery slope arguments, and the

ancient philosophical paradox of *Sorites* provides an example:

It is morally wrong to kill a sentient being, and a foetus at the time of birth (T) is a sentient being.

A foetus at one second (T-1) before the time of birth is also a sentient being, as the addition or

subtraction of one second cannot affect a being's sentience.

Therefore, a foetus at (T-2) is also a sentient being.

Therefore, a foetus at (T-n) is also a sentient being; a foetus at the moment of conception is a sentient

being.

The *Sorites* argument plays on the vagueness of the predicate "sentient", and the

inevitability of the logical inference of *modus ponens* to achieve paradox. The idea

that certain SSAs might be based on some kind of category boundary re-appraisal

mechanism has been articulated implicitly by many authors (Holtug, 1993; Lode,

1999; Rizzo & Whitman, 2003; see also Walton, 1992b, who distinguishes *sorites* as a

unique type of SSA). Indeed, the very notion that a slippery slope might exist between

an ostensibly positive proposal and a negative outcome directly implies an extension

process of some kind. When advances in gene therapy are discussed, the spectre of

Nazi eugenics is raised precisely because the concept of pro-social genetic

engineering is vague (Holtug, 1993), and membership of the category "acceptable

practice" is a dynamic and fluctuating process.

The message to unwary reasoners is that the majority of the concepts that pervade our

everyday argumentation are indeterminate. Because our everyday concepts lack

necessary and sufficient features and do not, as a consequence, have clear-cut

boundaries (for references see e.g., Pothos & Hahn, 2000), classification is heavily

dependent on the set of instances to which the category label has been applied.

Though very different accounts of the nature of conceptual structure exist, theorists

are agreed that there is a systematic relationship between the items that have been

classified as belonging to a category and subsequent classification behaviour. It is

fundamental to a wide range of current theories of conceptual structure that

encountering instances of the category at the category boundary will extend that

boundary for subsequent classifications. Furthermore there is a wealth of empirical

evidence consistent with these assumptions. In particular there are numerous

experimental demonstrations of so-called exemplar effects, that is, effects of exposure

to particular instances and their consequences for subsequent classification behaviour

(e.g., Nosofsky, 1986, 1988a, 1988b; Lamberts, 1995). For example, observing that a

dog that weighs 10kg is considered underweight invites the conclusion that a dog that

weighs 10.5kg is also underweight. With only the information that a 5kg dog is

underweight, and a 15kg dog is overweight, however, one might not be so compelled

to draw this conclusion[11].

There is, then, a feedback loop inherent in the classification of new data into an

existing category, whereby that classification also affects and alters the category itself

(see, e.g., Lakoff, 1987). In appropriate circumstances this extends the category

boundary in a way that could naturally give rise to slippery slope arguments (Hahn &

Oaksford, 2007a). This suggests that SSAs will have an empirical basis in many

cases; extending the cases that fall under a conceptually vague term will genuinely

facilitate future extensions. In other words, some slopes really are slippery.

---

[11] These examples of similarity-based categorisation are a specific instance of the more general process of similarity-based induction – see, e.g., Sloman (1993), or Osherson, Smith, Wilkie, Lopez & Shafir (1990).

However, this leaves open the question of whether or not people are also naturally aware of this in argument evaluation. Interestingly, exemplar effects in cognitive psychology have a broader counter-part in social psychology in the form of Social Judgment-Involvement theory (SJI) (e.g., Schwarz and Bless, 1992; Sherif, Sherif & Nebergall, 1965). According to this view, people's attitudes provide an interpretative context for incoming persuasive messages. If the position articulated in this message falls close enough to their own (i.e. falls within their 'latitude of acceptance') the message position is "*assimilated*". Message positions that are more distant (i.e. fall into the 'latitude of rejection'), by contrast, will be perceived as *more* dissimilar and distant than they truly are and will be rejected, leaving attitudes unchanged. In other words, the kinds of similarity effects that underpin category extension by exemplar have a counterpart in theories of attitude change and persuasion. This suggests that participants' evaluations of SSAs should also be sensitive to such effects. The next experiments set out to test directly whether parallel effects could be obtained in a categorization and an argument evaluation task that manipulated exemplar similarity.

Experiment 7 was designed to investigate the link between category boundary re-appraisal and slippery slope arguments using a uni-dimensional, quantitative category and numerically defined exemplars. If SSAs have an objective basis in category expansion driven by exemplar effects, there should be agreement between the perception of an SSA's strength and corresponding categorization decisions, given identical data to evaluate.

5.3   Experiment 7

5.3.1   Method

Participants

60 undergraduate psychology students from Cardiff University participated in the experiment in exchange for course credit.

Design, Materials and Procedure

All participants were presented with the following cover story:

In Finland, some rural locations are designated by the Government as areas of Outstanding Natural Beauty. If an area is identified as being of Outstanding Natural Beauty, no development is permitted on it. At the same time as preserving unique natural habitats, however, the Finnish Government must provide housing for its growing population. Land is designated as being of Outstanding Natural Beauty if it contains an unusually high number of large animal species.

Having read the cover story, all participants were presented with the following information about the decisions already made by the Finnish Government:

Location A

South Pernothea is home to **114** species of large animals.

Decision: Not eligible for Area of Outstanding Natural Beauty status.

Location B:

Reklan is home to **149** large animal species.

Decision: Not Eligible for Area of Outstanding Natural Beauty status.

Location C:

Grenadia is home to **259** types of large animals.

Decision: Eligible for Area of Outstanding Natural Beauty status.

Location D:

Scarathon is home to **224** species of large animals.

Decision: Eligible for Area of Outstanding Natural Beauty status.

All participants were then presented with two further locations (*Location I* and

Location X) that had not yet been decided by the Finnish government. The similarity

of *Location I* to *Location X* was manipulated by altering the value of the number of

large species in *Location I*. In one group, *Location I* was *similar* to *Location X*:

Two further cases are currently being considered by the Finnish government and the Finnish Housing

Association, the details of which are as follows:

Location I:

Aunskop is home to **194** species of large animals.

Location X:

Sellenfeld is home to **179** species of large animals.

In the second group, *Location I* was *dissimilar* to *Location X*:

Location I:

Aunskop is home to **218** species of large animals.

Location X:

Sellenfeld is home to **179** species of large animals.

The number of species of large animals was manipulated between the groups of participants, and two experimental measures (either a categorization decision or a rating of argument strength) were recorded, creating a total of four experimental groups.

In the categorization groups, participants were asked to make a categorization decision of their own, based on the information they had just read – i.e. whether *Location X* was eligible for *Outstanding Natural Beauty* status. The number of large animal species in *Location I* (i.e. the value of the *Location I* exemplar) was expected to differentially influence categorization decisions between the two groups. Participants who were told that *Location I* contained *194* animal species should be more likely to categorize *Location X*, with its 179 species, as eligible for *Outstanding Natural Beauty* status, as *Location X* is most similar to *Location I*. When *Location I* contained 218 animal species, however, *Location X* was more similar to the ineligible locations – and so participants should be less likely to categorise *Location I* and *Location X* together.

Participants in the argument conditions, by contrast, rated arguments based on the same materials. Argument strength was assessed by presenting *Location I* as part of a slippery slope argument. Participants were told that while the Finnish Housing Association was not too concerned about *Location I* being awarded *Outstanding Natural Beauty* status, this would lead to a further location (*Location X*) also receiving *Outstanding Natural Beauty* status, which the Finnish Housing Association viewed as problematic. It was predicted that participants who viewed this argument when *Location I* contained *194* animals would provide a higher rating of argument strength,

as they should perceive *Location X* as sufficiently close to the category boundary

defined by *Location I*, and therefore vulnerable to a slippery slope style re-appraisal

(mirroring the exemplar effect predicted in the categorization groups).

5.3.2   Results and Discussion

The yes/no data obtained from the categorization groups were analysed using a ranked

sign test. Participants who had been told that *Location I* contained *194* animals

categorized the new location as deserving of *Outstanding Natural Beauty* status on 11

of 15 occasions. Participants who had been told that *Location I* contained *218* animals

categorized the new location as deserving of *Outstanding Natural Beauty* status on 0

of 15 occasions. This difference was significant at p <.01.

The argument rating data were analysed using a t-test. Participants who had been told

that *Location I* contained *194* animals rated the arguments as significantly more

convincing (M = 4, SD = 2.1) than participants who had been told that *Location I*

contained *218* animals (M = 2.6, SD= 1.5), t (28) = 2.08, p < .05. These results

provide empirical support for the philosophical analysis of slippery slope arguments

by authors such as Govier (1982) and Volokh (2003) by demonstrating, in a tightly

coupled design, how slippery slopes may rest on a category boundary extension

process.

5.4 Experiment 8

Crucial to the analysis of the SSA that has been developed in this thesis is the idea that an undesirable outcome may be *re-evaluated* as acceptable following the implementation of an initial proposal. Anti-drugs campaigners suggest, for example, that the legalisation of cocaine might not seem such an abhorrent proposal if the legalisation of cannabis were to go ahead. In Chapter 4, I claimed that a defining characteristic of SSAs is that the predicted outcome may be *re-evaluated* – not simply evaluated differently by different groups of people. While group level re-evaluative processes are clearly an important component in many contexts where SSAs are to be found, it seems important to determine that the process of category boundary re-appraisal can also be demonstrated on an individual level. Experiment 8 was therefore designed to replicate the categorization data obtained in Experiment 7 using a within-participants design.

5.4.1 Method

Participants

40 undergraduate psychology students from Cardiff University participated in the experiment in exchange for course credit. Experimental conditions were manipulated within-participants, and so all participants completed the same experimental tasks.

Design

Experiment 8 was designed to provide a within-participants replication of the between-participants categorization effect observed in Experiment 7. Participants

173

were presented with a list of locations that had already been assessed by the Finnish

government (identical to the list used in Experiment 7). The independent variable was

manipulated by presenting two exemplar locations (Location I and Location V)

sequentially, differing in their numerical value. The dependent measure was assessed

across two categorization decisions, following the presentation of each exemplar

location. Effectively, participants in Experiment 8 completed both the experimental

tasks that had been split between the groups of categorization participants in

Experiment 7. The first categorization decision was made on the basis that Location I

(the first exemplar), containing 218 species of large animal, was eligible for

Outstanding Natural Beauty status. It was predicted that, in line with the results

obtained in Experiment 7, most participants would perceive Location X (the

dependent location, containing 179 animal species) as sufficiently dissimilar to

Location I to be *undeserving* of Outstanding Natural Beauty status.

The second categorization decision was made following information about Location

V (the second exemplar) containing 194 species of large animal, which was also

eligible for Outstanding Natural Beauty status. It was predicted that, in line with the

results obtained in Experiment 7, a significant number of participants would perceive

a similarity between Location X and Location V, and re-appraise their categorization

decision – i.e. classify Location X as *deserving* of Outstanding Natural Beauty status.

Materials and Procedure

Each participant received an experimental booklet containing a brief description of

the fictitious scenario, a list of locations that had already been adjudicated, and the

dependent measure tasks (i.e. two categorization decisions). The booklet was

constructed so that participants would not see the additional information, or the second categorization task, until they had made their first decision.

5.4.2 Results and Discussion

The distribution of categorization responses is presented in Figure 5.1. A McNemar Change test of the difference between the distributions revealed a significant shift in categorization behaviour following the presentation of the second exemplar location, $p < .01$.
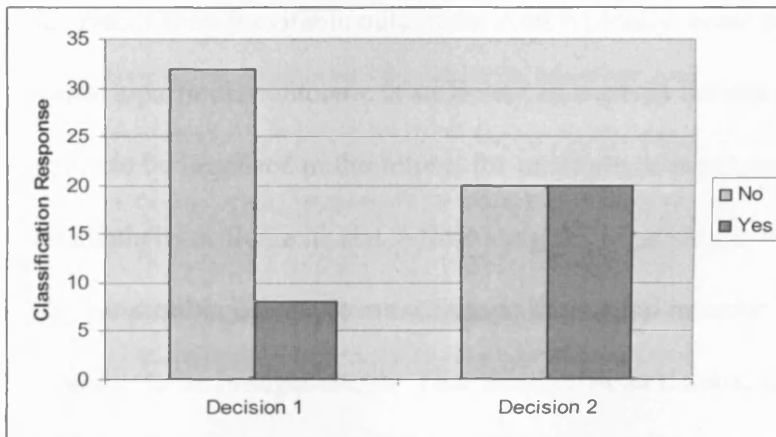
Figure 5.1: Distribution of categorization responses in Experiment 8.

The minority of participants who had already responded "yes" on the first categorization task all maintained this decision on the second categorization task. This demonstrates that people were not simply changing their mind in response to being asked to make the same decision twice. Excluding these participants from the statistical analysis (as it is not possible to extend a category to incorporate data it already includes), did not alter the finding of a significant difference across the two decisions ($p < .001$).

175

The reader may note that the distribution of YES/NO responses on the second classification task is equal – i.e. half the participants perceived the dependent location as *not* sufficiently similar to the exemplar location to warrant classifying them together. The focus of the experiment was, of course, on the relative change in response distribution, rather than the absolute magnitudes of the distributions themselves. However, finding that re-evaluation is probabilistic, rather than deterministic (i.e. it will occur for some individuals but not others), fits well with both the Bayesian account of SSA strength and an intuitive understanding of how slippery slopes are realised in practical contexts. Typically, SSAs warn of threats, dangers and risks, rather than inevitable outcomes. And typically, *some* change in attitudes towards a particular outcome is sufficient to warrant the use of an SSA. In order for cocaine to be legalised in the future, for example, it is not necessary that *all* individuals re-evaluate its status following the legalisation of cannabis. Rather, a *sufficient* number of people must change their mind in order for a qualitative shift in legislation to be brought about. That this 'sufficient' amount is typically undefined only serves to reinforce the idea that SSAs trade on the uncertainty of the future.

Taken together the results of experiments 7 and 8 suggest that the re-appraisal of category boundaries may provide an empirical mechanism underpinning the occurrence of (some) slippery slopes. Most natural categories lack clearly defined boundaries and are susceptible to exemplar-based modifications as documented here, and extensively elsewhere (e.g., Nosofsky, 1986, 1988a, 1988b; Lamberts, 1995). The probability of a predicted outcome following from an initial proposal appears to be directly related to the empirical similarity between the initial proposal and the predicted outcome. Furthermore, category boundary re-appraisal can occur not only

across groups of people, but within the same individual or group over time (for related

work on the flexible re-appraisal of information in social judgements, see Wegener &

Petty, 1995). In addition, the present experiments suggest that SSA evaluation is also

sensitive to exemplar manipulations, in line with Social Judgement-Involvement

theory and assimilation-contrast effects (Schwarz & Bless, 1992; Sherif, Sherif &

Nebergall, 1965; Stapel & Winkielman, 1998). In summary the experimental evidence

suggests a well-defined, objective basis, grounding subjective judgements of SSA

acceptability.

5.5   Experiment 9

Thus far the investigation of similarity-categorisation as a mechanism of the slippery

slope has been quantitative. This design was purposefully chosen to allow the effect

of the exemplar manipulation to be examined using uni-dimensional, numerical

stimuli. The experimental tasks provided participants with explicit numerical

information about exemplars, as well as an indication of the categories to which they

belong. This quantitative approach allowed the objective similarity between the

exemplar and existing category members to be controlled with some accuracy.

Generally, however, categories in the 'real world' are neither defined using explicit

numerical criteria, nor based on a single, quantifiable dimension. Correspondingly

SSAs are usually not about such categories. Instead SSAs typically relate to notions

such as personal freedoms (the legalisation of voluntary euthanasia) or civil rights (the

introduction of ID cards). Accordingly, it seems necessary to replicate the results of

the exemplar-based slippery slope experiments using materials that bear a closer resemblance to SSAs as they might appear in a natural context.

Participants were presented with some simple information about the legal system of a fictional location 'Sotherby Island'. Participants were informed that all crimes committed on Sotherby Island would now be punished by either less than 20 years in prison (SUB 20 crimes), or by more than 20 years in prison (20 PLUS crimes). Following the design of the previous experiments, an argumentation and a categorization measure were developed based on these materials.

5.5.1   Method

Participants

52 undergraduate students from Cardiff University participated in Experiment 9 in return for course credit.

Design

All participants were presented with the following information:

The inhabitants of Sotherby Island have decided to clear up their complex legal system once and for all, by creating a two-tier sentencing structure. They have voted for a strict division of sentences, such that some crimes are automatically punished by 20 years (or more) in prison (the 20 PLUS category), and all other crimes are automatically punished by less than 20 years in prison (the SUB 20 category). The islanders have already voted on the categorization of many crimes, some of which are shown below.

Offence: Burglary   Offence: Manslaughter

Decision: SUB 20 category   Decision: 20 PLUS category

Offence: Rape

Decision: 20 PLUS category

Offence: Arson

Decision: SUB 20 category

Offence: Murder

Decision: 20 PLUS category

All participants in the experiment completed both an argumentation *and* a

categorization measure, which were combined in the same set of materials.

Participants were presented with two further crimes, which had not yet been voted on

by the Sotherby Islanders. The qualitative exemplar manipulation was implemented

by altering the similarity of these two crimes to each other, creating two experimental

groups. Participants in the *similar* experimental condition ($n = 27$) were presented

with these two crimes:

Offence A: Assault in Possession of a Knife

Decision: ???

Offence B: Assault in Possession of a Gun

Decision: ???

Participants in the *dissimilar* experimental condition ($n= 25$) were presented with

these two crimes:

Offence A: Assault without a Weapon

Decision: ???

Offence B: Assault in Possession of a Gun

Decision: ???

As in previous experiments, an SSA was constructed stating that *if* **Offence A** is made a SUB 20 crime, *then* **Offence B** will also be made a SUB 20 crime, and that **Offence A** should therefore be made a 20 PLUS crime. Participants were asked to evaluate the strength of this argument, but in this experiment three measures of argument evaluation were obtained. Participants had to indicate on a scale from 0 (Unconvincing) – 10 (Very Convincing) how *convincing* they found the islanders argument, on a scale from 0 (Unpersuasive) – 10 (Very Persuasive) how *persuasive* they found the argument, and on a scale of 0 (Very Weak) – 10 (Very Strong) how *strong* they found the argument. These measures were intended to measure the same underlying construct – i.e. argument evaluation – although any differences that emerged would permit a more stringent analysis of how participants were interpreting the argument evaluation task.

Participants were then informed that the Sotherby Islanders *did* decide to make **Offence A** a SUB 20 crime. Their next task was then to decide whether **Offence B** should also be made a SUB 20 offence. Participants were also asked to indicate on a continuous scale from 0 (Definitely Not) – 10 (Definitely) whether they thought **Offence B** should be made a SUB 20 crime.

In an attempt to avoid the potential effects of prior beliefs about which crimes should and should not receive lengthily jail sentences in the 'real world', participants were explicitly instructed to make their ratings of argument strength and categorization decisions based *only* on the information they had seen in the experiment. Specifically, participants were requested not to consider whether (in 'real life') it would be morally

right or wrong to place any particular crime in any particular category, but simply to make their assessments based on the information in the experiment.

5.5.2 Results and Discussion

Argumentation Data

A reliability coefficient (Cronbach's Alpha) for the three measures of SSA evaluation was computed to be .89. This suggests that the three measures can be interpreted as measuring the same underlying construct (i.e., SSA evaluation). Responses on the three measures of SSA evaluation were therefore combined in order to create a single measure of argument evaluation.

A *t*-test was conducted to examine the effect of the exemplar manipulation. Ratings of argument evaluation were significantly higher in the similar condition ($M = 5.02$, $SD = 1.8$) than in the dissimilar condition ($M = 2.66$, $SD = 1.7$); $t(50) = 4.77$, $p < .001$.

Categorization Data

Participants in the similar experimental condition gave significantly higher ratings ($M = 5.18$, $SD = 2.68$) of whether **Offence B** should be placed in the same category of crimes as **Offence A** than participants in the dissimilar condition ($M = 3.8$, $SD = 2.25$). This difference was statistically significant, $t(50) = 2.01$, $p < .05$.

The results of Experiment 9 show differential SSA evaluations and categorization preferences based on a qualitative exemplar manipulation, analogous to the existing data (experiments 7 and 8) showing the same effect with a quantitative exemplar

181

manipulation. To this extent, the claim that similarity-based categorisation is a plausible psychological mechanism on which evaluations of SSA strength may be based is strengthened. The data in the current experiment demonstrate that exemplar based similarity judgments in the context of SSA evaluation do not require explicit numerical information in order to be observed. Rather, the mechanism of exemplar similarity seems generalisable to more qualitative contexts where SSAs are typically found.

5.6 General Discussion

The three experiments reported in this chapter demonstrate how and why there can be an *objective*, non-zero probability that the re-evaluation or 'slippage' on which SSAs are predicated can, in fact occur. In other words, some slopes really are slippery, because their beginning and end are similar. Experiments 7 and 8 showed this by linking experimentally categorization and slippery slope argument acceptability. Exemplar effects provide the kind of empirical mechanism that the fear of outcome re-evaluation inherent in slippery slope arguments requires. Having demonstrated this using quantitative experimental stimuli, Experiment 9 extended this account using experimental materials that could only be categorized in terms of qualitative similarities. This suggests that an exemplar-similarity account applies equally to a wide range of possible categories – and therefore a wide range of real world contexts. Crucially, these three experiments not only demonstrate exemplar effects (which are well known) but suggest that people understand and naturally take into account such effects when making judgments about SSAs – in precisely the way that theories of assimilation and contrast effects (Schwarz, 1996; Schwarz & Bless, 1992) and the

Social Judgement Involvement theory of persuasion (Sherif, Sherif & Nebergall, 1965) suggest that they should. Not only do exemplar effects in the context of conceptual vagueness provide a widely applicable underlying mechanism for real-world slippery slopes that underwrites their objective rationality, people seem naturally sensitive to this mechanism in their own subjective evaluations of SSAs: Their ratings of arguments strength mirror their categorisation judgements, both within and between participants. Finally, the data clearly show how perceptions of argument strength are influenced by the facts that form the content of an argument, not simply by procedural rules of discourse as pragma-dialectical theories assume.

Of course, similarity-based categorisation is not the *only* mechanism on which SSAs may be predicated. Walton (1992b), for example, has outlined an SSA classification system that distinguishes 'precedent' SSAs (whereby each step on the slippery slope sets a new precedent), and 'causal' SSAs (whereby each step causes the next)[12]. Clearly, while a mechanism based on category boundary re-appraisal can deal comfortably with precedent SSAs – after all, precedents are simply category markers – causal SSAs are something distinct. From a psychological point of view, however, causal SSAs might be thought of as less mysterious, as there is no need to posit any additional mechanism: Causal SSAs proceed because they embody a (perceived) chain reaction, not because of re-evaluation. Whether or not a particular causal process will actually bring about the undesirable outcome at the end of a slippery

---

[12] In fact, Walton (1992b) distinguishes between three types of SSA, separating the 'sorites' SSA from the precedent and causal arguments where "reasoning with a key vague term is evidently the source of the problem" (p2). To the extent that vague category boundaries facilitate category boundary re-appraisal, however, it seems unnecessary to consider the sorites separately from the precedent slope for current purposes. Seemingly essential philosophical distinctions do not necessarily map precisely on to psychological mechanisms – and while this does not undermine the usefulness of Walton's taxonomy, it is also clear that empirical, psychological work should not be bound by it.

slope could also be straightforwardly re-described as a probabilistic judgement – how *likely* is A to cause B to happen?

Volokh (2003) has identified several other ways in which SSAs might genuinely be slippery. Some of these mechanisms (including the notion of 'multi-peaked voter preferences') would be amenable to a psychological analysis, and I discuss these in detail in the following chapter as a possible direction for future research. Given, however, that many of the most famous slippery slopes (e.g., euthanasia, abortion, drug classification, civil liberties) seem to involve re-appraising category boundaries in one way or another, the linking of category judgements and argument strength seems an appropriate way with which to conclude my examination of SSA evaluation.

5.7   Chapter Summary

In this chapter I have reported three experiments demonstrating that when an SSA seems probabilistically strong, this may be because some slopes *really are slippery*. By linking similarity-based categorisation judgements, and ratings of argument strength, I have shown parallel effects of manipulating exemplar similarity: When the initial action and predicted outcome of an SSA are more similar, they are more likely to be categorised together, and comprise a probabilistically more compelling argument. The main contribution of this chapter is in showing that SSAs are not only subjectively rational – their subjective strength may be based on an objective measure of exemplar similarity.

The experiments reported in this chapter conclude the empirical work conducted in this thesis. In the next chapter, I summarise the work I have presented, discuss some outstanding questions raised by it, and outline some potential directions for future research.

Chapter 6 - Summary, Outstanding Questions and Future Directions

6.1   Chapter Overview

The purpose of this chapter is to draw together the work that I have presented in this thesis, but also to identify the outstanding questions that my research raises. I will try to provide some overall conclusions, identifying the central contributions of this thesis to furthering knowledge about how people evaluate informal arguments.

6.2   Overview of Thesis

The goal of this thesis was to contribute towards answering a simple, longstanding and important question: How do people evaluate informal arguments?

In attempting to answer this question I have drawn on empirical findings relating to reasoning, persuasion and attitude change, categorisation and decision-making, as well as theoretical analyses from disciplines as diverse as social and cognitive psychology, legal philosophy, epistemology and science communication. The theoretical framework underpinning the work presented in this thesis is Bayesian probability theory, which treats arguments as claims (or hypotheses) that are backed by evidence, and people's degrees of belief in hypotheses as probabilities. On the Bayesian account of informal argumentation, an argument is strong to the extent that it provides evidence in support of a hypothesis (Hahn & Oaksford, 2006a, 2007a). Bayes' Theorem provides a rational method for incorporating new evidence into existing beliefs. The Bayesian approach provides a normative, quantitative measure of

argument strength, and permits arguments to be analysed and experimentally manipulated according to their individual content. Expanding on the Bayesian analyses of argumentation developed in Hahn and Oaksford (2006a, 2007a), I have conducted an in-depth investigation of how people evaluate informal arguments, in two distinct but related contexts: Socio-scientific arguments and SSAs.

The single biggest advantage of bringing the Bayesian framework to bear on the issue of informal argument evaluation is that it permits a broad range of questions about argument strength to be studied using a very minimal set of theoretical assumptions – specifically, that people's beliefs can be described probabilistically, and that it is rational for people to calibrate their degrees of belief to the probability calculus. The Bayesian claim to normativity is based on combining the self-evident, mathematical axioms of probability theory with a basic economic rationale (i.e. the Dutch Book Argument). I have argued that as normative theories go, the Bayesian framework provides sound normative guidance for argumentation (Chapter 2). Using the simple assumptions of the Bayesian approach, it is possible to conduct empirical research that bears on contemporary problems such as the communication of science (Chapter 3), and on longstanding philosophical puzzles such as the slippery slope argument (Chapters 4 & 5). People's evaluations of a variety of informal arguments can be compared to Bayesian predictions, and arguments on different topics compared to each other. While these achievements may sound straightforward, the study of informal argumentation has been dominated by a historical dependency on the laws of formal logic, and more recently by an increasing focus on procedural and dialectical rules. I have argued that neither of these theoretical approaches provides the level of

analysis necessary for a theory of argument strength – that is, analysis at the level of argument *content* (see Chapter 1).

To a certain extent the experiments in this thesis were simply a vehicle for posing general questions about argument strength that are equally applicable to other types of informal argument. Of course, different types of argument have idiosyncratic characteristics, and in the experimental work I have presented I have tried to tailor the theoretical analyses to these features as much as possible. For example, my examination of existing work on scientific argument evaluation required a consideration of Toulmin (1958) and the dialectical approach to argumentation, whereas SSAs link more closely to social psychological literature on persuasion and attitude change. At root, however, the Bayesian approach is a general one and is not linked to a particular argument type, topic, or dialectical setting. This generality is a significant advantage for a theoretical framework designed to capture a notion as broad as argument strength. The breadth of the work presented in this thesis is a direct consequence of the generality of the Bayesian approach.

The choice of socio-scientific arguments and SSAs as the focus of the experimental work was chiefly because of the analytical opportunities they provided. The study of the evaluation of scientific arguments is a relatively recent concern, but one that is becoming increasingly pressing in the light of challenges such as the effective communication of climate change research. However, despite wide-ranging academic interest in the communication of science, a psychological understanding of how people evaluate simple scientific arguments has not been forthcoming. While a complete understanding of the communication of scientific messages cannot be

obtained using only an experimental method, such an approach permits questions to be posed systematically and with a high degree of control. Given that the goal of most research on science communication is to 'improve' the public understanding of science, it seems essential to understand whether, on a basic, psychological level, there is anything 'special' about the evaluation of scientific arguments.

SSAs, on the other hand, are a longstanding puzzle in philosophy but have not received an empirical treatment. And because they combine both probabilities and utilities, they offer a rich vein for a psychological analysis of the factors that determine their effectiveness. By pursuing an in-depth examination of SSAs, I have been able to offer some empirical answers to questions about this so-called fallacy that have historically resided solely in the philosophical domain.

6.3 Outstanding Questions and Remaining Issues

In what follows, I will identify what seem to be the outstanding questions raised by the analyses in this thesis. Rather than summarise each chapter individually again, however, I will split my evaluation and discussion into five different sections, which I consider to be the key areas in which the present work could be extended or improved.

6.3.1 Bayesian Normativity and Experimental Pragmatics

Chapter 2 was designed to provide justification for the claim that the Bayesian approach provides a *normative* theory of argument strength. By looking outside of the

reasoning and argumentation literatures for guidance on what legal philosophers and epistemologists have said about norms and normativity, I attempted to outline the features that a good normative theory should possess. However, the issues surrounding normativity are complex, and the question of what makes something normative is one that has occupied philosophers for centuries (see, e.g., Bishop & Trout, 2005). I therefore restricted my analyses to a simpler question: What makes a good normative theory of *argument strength*? The case for Bayesian probability theory as a normative framework for evaluating informal arguments is, I think, a strong one. However, I cannot claim to have provided a definitive answer to my question – due, in part, to the lack of philosophical consensus over what makes a good normative theory in *general*.

One crucial aspect of the debate over whether the Bayesian approach provides sensible normative principles for reasoning and belief revision is the extent to which beliefs, and networks of beliefs, are localised. If, as is increasingly claimed by proponents of Bayes' nets, many of our beliefs can be effectively isolated from causally unrelated evidence (see, e.g., Pearl 1988), then the problems of computational complexity that would seem to render the task of maintaining probabilistic belief consistency impossible, may be substantially mitigated. Encouragingly, this is a question that is potentially testable using empirical methods. For example, recent research has utilised the framework of causal Bayes' nets to examine the way in which people evaluate complex legal evidence (Lagnado & Harvey, in press).

However, as one difficult question is tackled, another equally difficult one arises: When, if ever, can evidence be said to be 'causally unrelated'? While the development of Bayes' nets allows the modelling of localised belief networks, any such model is necessarily dependent on certain assumptions about whether or not evidence is causally related to a particular hypothesis. It is one thing to show that evidence does not impact on causally unrelated beliefs (suggesting that probabilistic consistency may not be as intractable as it appears to be); it is quite another to explain *why* some evidence is causally relevant, and some not.

In Chapter 2, I explicitly avoided attempting to explain any of the (substantial amount of) experimental data that suggests that people are often not Bayesian in their evaluation of hypotheses and evidence (see, e.g., Gigerenzer & Goldstein, 1996; Nisbett & Ross, 1980; Tversky & Kahneman, 1983). The many reasons why people may not, in experiments or their everyday lives, reason in accordance with Bayesian norms have been discussed in great detail elsewhere (Hilton, 1995; Oaksford & Chater, 2007; Stanovich, 1999). In addition, the ontological separation between the normative and the descriptive means that a norm's integrity does not necessarily hinge on descriptive data. In any case, the experiments in this thesis actually provide broadly consistent support for the claim that people are, or *can be* Bayesian in their evaluation of informal arguments and evidence. Thus on the current evidence the Bayesian approach seems to provide a fairly good descriptive account of informal argument evaluation, as well as sound normative guidance. But the continuing fascination with normative questions among reasoning researchers suggests that possessing the appropriate normative model is essential for accurate empirical conclusions to be drawn.

In fact, the normative construal of an experimental task can have wide-ranging

implications – a theoretical perspective that is often referred to as 'experimental

pragmatics' (Hilton, 1995: Noveck & Sperber, 2004; Schwarz, 1996; Stanovich &

West, 2000). The key insight is that in order to be able to accurately understand

behaviour in an experiment, it is vitally important to have a complete understanding

of what the *participants* in the experiment think they are doing, in case it differs from

what the *experimenters* think they are doing. Yet in many psychological studies of

reasoning the routine assumption is that participants' representation of the

experimental task matches that of the experimenter. Standard normative models tend

to consider only the specific content of experimental tasks, rather than the broader

context in which these tasks are completed (i.e. the experiment as a social interaction

between experimenter and participant).

Increasingly, some psychologists studying reasoning have been willing to consider the

pragmatics of the experimental setting when formulating normative models. These

researchers have based their analyses of reasoning behaviour on the Gricean notion of

conversational implicature: Information that is not contained in the literal content of

an utterance, but that can be implied from the context in which it is given (Grice,

1975). Grice proposed that people strive to adhere to certain *maxims* of conversation,

in particular the maxim of co-cooperativeness. Speakers endeavour to be cooperative,

but also to be relevant (the maxim of relevance), to be concise but not unnecessarily

so (the maxim of quantity) and to be accurate (the maxim of quality). By

understanding and applying these maxims, interlocutors can extract much more

information from an utterance than is contained in its literal content. That reasoners

strive to obey these Gricean principles is taken to be self-evident in the fields of

linguistics and pragmatics[13]. In psychological, empirical studies, however, the notion of conversational implicature has typically been overlooked.

Arguably, many of the most famous 'errors' committed by participants in reasoning experiments may be recast as pragmatically reasonable responses to communicatively complex problems (that have been poorly specified by the experimenter – Hilton, 1995). For example, a long standing 'bias' in the cognitive psychology literature is that participants use non-diagnostic information to classify items. Typically this has been taken as evidence that people's ability to categorize is not normative – they use irrelevant information in experimental tasks, and take longer than necessary to group items together (see, e.g., Nisbett, Zukier & Lemley, 1981). Hilton observed, however, that the design of these experiments violates a fundamental Gricean assumption – that co-operative communicators will not provide too much or too little information. Participants naturally assume that the same rules will apply in an experiment, yet are routinely provided with non-diagnostic information. Assuming that if it is included in the experiment it must have some relevance to the task, participants attempt to use all the information they have – only to be accused of irrational behaviour by psychologists.

In Chapter 2 I suggested that the guarantee of epistemic consistency that the Dutch Book Argument provides makes it unlikely that people would *not*, in general, wish to adhere to Bayesian norms for argumentation. The lesson from the experimental pragmatics literature, however, is that experimenter and participant do not always

---

[13] Gricean maxims are, of course, a basis for pragma-dialectical norms of argument acceptability, discussed in Chapter 1 and Chapter 2 (van Eemeren & Grootendorst, 2004). While I have suggested that they are insufficient for a content-level analysis of argument strength, Gricean conversational maxims provide an essential framework for assessing communicative intent.   ╱

share the same representation of an experimental task. On the one hand there is no reason to suggest that the interpretation of the experiments reported in this thesis differed between participant and experimenter: Participants' responses typically matched Bayesian predictions to a significant degree, and because the responses were *subjective* evaluations of arguments and evidence, any discrepancy between participants' construal of the task and normative predictions would have been obvious in the experimental data. On the other hand, however, the extent to which participants would agree that Bayesian norms for argumentation are ones that they would (ideally) subscribe to is an empirical question. One profitable way of extending the normative analysis presented in Chapter 2 would be to take Bayesian norms for argumentation and belief revision, and ask people to evaluate them. This question is distinct from establishing whether, in practice, people *actually follow* Bayesian norms: The issue at stake here would be whether people agree that they are the norms that *should* be followed.

6.3.2 Communicating Science: Communicating to whom?

Much of the work that has been conducted into the public understanding of science assumes that there are problems with the communication of science – certainly, the explicit goal of science educators is to improve scientific literacy, and the response of the public to messages about, for example, climate change, suggests that effectively communicating science is a complex task. However, on the Bayesian account of argument strength, scientific arguments are just arguments that *happen* to be about science – and the same factors that determine the strength of non-scientific arguments should also influence the evaluation of scientific arguments (indeed, historically,

194

Bayesian probability theory was first applied to formal scientific reasoning – see

Howson & Urbach, 1996).

The experiments reported in Chapter 3 provide an initial, experimental insight into

how people evaluate scientific arguments and evidence. By utilising an experimental

method, a degree of control was obtained such that individual factors like source

reliability and evidential coherence could be manipulated accurately. And while

science may not typically be communicated to the public in such a well-regulated

way, it is also true that most people's knowledge of socio-scientific topics is gained

through the media, where short, carefully packaged representations of science are the

norm. Given that these condensed arguments about scientific issues are prevalent in

the public domain, and considering how disparate the existing literature on scientific

argument evaluation is, the Bayesian approach has substantial heuristic value in

permitting crucial questions about scientific arguments to be posed at all.

However, the question of how people evaluate socio-scientific messages is a

fundamentally applied one – that is, the communication of science to different groups

of people in the 'real world' is likely to differ in important ways from such a tightly

controlled experimental analysis. Many of the factors that might be expected to

influence the communication of scientific messages (e.g., self-interest, or the

perceived status of message giver) were intentionally excluded from the experiments

in Chapter 3. One way of extending the experiments would be to conduct replications

of the experiments with key factors such as these measured or manipulated.

Moreover, all the participants in experiments 1a-1d were undergraduate students, and as such the more pressing need is to extend these basic experimental findings using different populations that might reasonably be expected to differ in their evaluation of scientific arguments – practising scientists, policy makers, or members of the public that are more representative of the general population than university students. Of course, the problem of experimental samples that poorly represent the general population is one endemic to much academic psychology. But for many research programmes there is no reason to suppose that the response of an undergraduate will differ from that of a middle-aged businessman. In the case of the public understanding of science, however, there are grounds for positing important differences in the way that different groups of the population think about and evaluate science.

For example, emerging research on how people use different styles of discourse to 'tell the story' of climate change suggests that there are distinct response to messages about environmental issues (Segnit & Ereaut, 2007) that vary according to demographic characteristics. Segnit and Ereaut identified linguistic repertoires (e.g. 'reluctant acceptance') that are associated with different political standpoints. This suggests that what makes a strong and compelling socio-scientific argument to one group of people may not be effective with another group. Of course this feature of argumentation, known as audience relativity (see, e.g., Perelman & Olbrects-Tyteca, 1969; van Eemeren & Grootendorst, 2004), is well studied in the social psychological literature on persuasion (see, e.g., Kaplan, 1971; Petty, Cacioppo & Goldman, 1981). And because the Bayesian approach deals with *subjective* probabilities, it can straightforwardly incorporate differences in prior beliefs into predictions about belief revision and argument evaluation. However, it should still be the case that differences

between groups in terms of argument evaluation are predicted by Bayesian theory –

so while a climate sceptic might distrust a scientific consensus this should be

detectable in the judgements of the reliability of scientists as sources of information.

The Bayesian approach has no difficulty in incorporating subjective differences in

prior belief into predictions about argument strength and therefore offers an ideal

framework with which to expand the scope of the experiments reported in Chapter 3.

6.3.3   Alternative SSA Mechanisms

In Chapter 5 I reported three experiments demonstrating that subjective evaluations of

SSA strength may be predicated on objective, measurable differences in the similarity

between the 'top' and the 'bottom' of a slippery slope. Specifically, the more similar

the beginning and the end of the slope are, the stronger the SSA will be. Furthermore

people's similarity-based categorisation judgments change in the light of new

information – suggesting that some slopes *really are slippery*. Many of the most

famous examples of SSAs seem to involve a category boundary re-appraisal of some

kind, including arguments relating to drug classification, civil liberties, abortion and

euthanasia. But category boundary re-appraisal is not the only way that a slope might

genuinely be slippery.

Volokh (2003), discussing slippery slopes in the legal domain, outlined a number of

mechanisms on which SSAs might be predicated including something known as

'multi-peaked voter preferences'. According to Volokh, in many legal debates the

public can be divided into three groups: Traditionalists, who don't want to change the

law because they prefer the current position 'A'; Moderates, who want to change the

law a bit to position 'B'; and Radicals, who want to go all the way to position 'C'.

Volokh uses the example of CCTV cameras to illustrate his point, where position 'A'

means no CCTV cameras on the street corner, position 'B' means CCTV cameras but

no archiving and face recognition, and position 'C' means CCTV cameras with

archiving and face recognition. Typically, 'single-peaked preferences' can be

assumed – that is, both traditionalists and radicals would prefer position 'B' to the

extreme on the other side. Problems can arise, however, if the single-peaked

preference assumption does not hold:

"...(S)ay instead that some people prefer A best of all (they'd rather have no cameras,

because they think installing cameras costs too much), but if cameras were installed

they would think that position C (archiving and face recognition) is better than B (no

archiving and no face recognition): 'If we spend the money for the cameras,', they

reason, 'we might as get the most bang for the buck.' This is a multi-peaked

preference – these people like A *least*, preferring either extreme over the middle"

(Volokh, 2003, p1049).

The implication of this hypothetical situation is that the supposedly moderate mid-

point is actually very unstable – because both traditionalists and radicals prefer it the

least. It is not difficult to see how unstable middle ground would provide precisely the

impetus required for a slippery slope.

Volokh's analysis poses questions that could be tackled empirically. For example, one

could ask participants to indicate their positions on political issues, divide them into

three groups (i.e. traditionalists/moderates/radicals) based on their responses, and then

ask them to 'vote' on their preferred outcomes. For topics where the moderate

position was consistently preferred the least, SSAs should be perceived as stronger –

because slippage really is more likely. Establishing the empirical validity of multi-

peaked preferences would provide another insight into the psychology of SSAs, and

would be a fascinating avenue for potential future research.

6.3.4   Capturing Argument Strength

Throughout the empirical work presented in this thesis I have referred to the Bayesian

approach as providing a measure of *argument strength*. In accordance with the

Bayesian definition of an argument as "claim-plus-evidence", I provided participants

with a claim (e.g. that 'drug A is safe') and some evidence that bears on this claim

(e.g. 'ten experiments have found no side effects'). I have typically defined the

strength of an argument as the rating that participants assigned to a particular claim –

that is, their probabilistic degree of belief that the claim is true. However, there are a

number of subtly different ways to elicit a judgment of argument strength depending

primarily on the particular question that one asks. Consider the following two ways of

measuring the strength of an argument regarding the safety of 'Drug A':

(1)   What is the chance that Drug A is safe, given that ten experiments have found no side

effects?

(2)   How much do ten experiments finding no side effects confirm that Drug A is safe?

The first question asks for a judgment of the truth of the hypothesis given the

evidence, and is representative of the way in which I have elicited judgments of

argument strength from participants throughout the work presented in this thesis – that is, I have asked participants to indicate how convinced they are of the truth of the hypothesis. This quantity can be described as their posterior degree of belief, or their ultimate degree of conviction: An argument is strong to the extent that it convinces someone of the claim it seeks to support.

The second question, however, asks how much the evidence confirms the hypothesis – that is, it is a measure of how much *more* likely the hypothesis is to be true following the provision of the evidence, than *before* the evidence had been provided. This question can be described as a measure of *confirmation*, and there is currently a good deal of debate over which, of several competing measures of Bayesian confirmation, provides the most accurate measure of evidential support (see Tentori, Crupi, Bonini & Osherson, 2007). I do not seek to resolve or discuss here which measure of confirmation is ultimately preferable – it matters for the present context only that such measures can clearly be derived. Suffice to say, a Bayesian measure of confirmation must provide some way of capturing the discrepancy between prior and posterior belief.

Measures of confirmation are likely to be important for a complete Bayesian theory of argument strength. This is because the strength of an argument is influenced in part by the degree of belief an individual *already has* in the hypothesis (i.e., their prior), so that an argument will be more convincing if one is already positively inclined toward the claim in question. To illustrate this point, consider an individual who is already very convinced that Drug A is safe. This could be for any number of reasons –

perhaps they have taken medications made by the company that makes Drug A before. Imagine that they then receive an argument regarding Drug A, which reads:

"Drug A is safe because Saturday follows Friday"

Clearly, the argument is weak, as it provides no evidence whatsoever about the safety of Drug A. But if one were to ask the individual how convinced they were that Drug A was safe, one would still obtain a high rating of convincingness – because they are already maximally convinced that the Drug is safe. If one were to ask how much the fact that Saturday follows Friday *confirms* that Drug A is safe, however, an entirely different answer would be obtained. Ultimate degree of conviction does not *necessarily* provide useful information about the strength of an argument – as it is 'contaminated' by prior beliefs.

The reader might question whether the experiments reported in this thesis really get to the heart of the argument strength question, given that the Bayesian framework naturally emphasises *changes* in degrees of belief. Indeed, this is an orientation that fits well with argumentation's fundamental goal of seeking to convince someone, given that it is someone who does not already share our convictions that we typically argue with. In the work presented in this thesis, however, this problem arguably does not really arise.

Firstly, two experiments (1c; 6) did explicitly measure belief change, and experiments 8 and 9 demonstrated category boundary re-appraisal on an individual level. Secondly, the majority of the materials used to elicit judgments of argument strength

utilised fictional topics[14]. Participants could not have had *any* prior beliefs in the truth

or falsity of the claims provided in the arguments, other than the evidence provided –

the only evidence that participants had to evaluate the truth of the claims was

provided in the experiments themselves. A fundamental advantage of using fictional

materials and an experimental methodology to study argumentation is that variables

such as the amount of evidence, or source reliability can be precisely and accurately

controlled – without the unknown influence of prior beliefs.

Crucially, however, if participants had been incorporating pre-existing beliefs into

their experimental judgements it would not have artificially enhanced the data in this

thesis – if anything it would have dampened the effect of any experimental

manipulation. If participants were responding with only their pre-existing prior

beliefs, one would not expect to observe such consistent and reliable differences in

experimentally induced ratings of argument strength. So, while measures of change

and confirmation are ultimately important components of a complete Bayesian theory

of argument strength, they speak more to the wider issue of obtaining a 'pure'

measure of argument strength, than they do to the integrity of the experimental data in

this thesis.

One profitable way of expanding the current work would be to test for correlations

between as many different putative measures of argument strength and confirmation

as possible – while some may be interchangeable, others may isolate distinct aspects

---

[14] Some experimental materials were based on real-world events (e.g., anthropogenic climate change,

or the introduction of ID cards in the UK). However, participants were always instructed to use only

the information provided in the experiment to make their judgments.

of argument evaluation. The value of experimental responses will fundamentally depend on the usefulness of the questions that the experimenter asks (see Nelson, 2005, 2008): It is therefore essential that the right questions are asked. Developing the work in this thesis may require a more refined definition of argument strength – and an acknowledgement that different questions may be more appropriate or informative depending on the context. However, it is only once exploratory work establishes that the Bayesian framework *does* in fact provide a profitable way of studying judgments of informal argument strength that more finely tuned questions can be posed. Identifying and distinguishing different components of the broader notion of 'argument strength' will be a crucial tool for moving this type of experimental research forward.

In addition to distinguishing different measures of argument strength in future work, the notion of 'source reliability' could also be subjected to a closer examination. In several of the experiments reported in this thesis, I have manipulated the perceived reliability of the source in the arguments that participants received. In Experiment 1a, for example, source reliability was manipulated by informing participants that either a respected academic journal (a reliable source) or an internet blog (an unreliable source) had reported evidence that a pharmaceutical drug was safe. Bayes' Theorem predicts that the perceived reliability of the source reporting the evidence should have a systematic impact on the strength of the argument. The data in this thesis support that prediction.

However, there is more than one way that the notion of reliability can impact on the strength of evidence. As well as the perceived reliability of the source *reporting* the

evidence (i.e. a respected academic journal vs. an internet blog), the evidence itself may be more or less reliable. That is, the tests that the sources report may be more or less reliable and this should also impact on how compelling the evidence is. So while a respected academic journal may be a highly reliable source, the evidence it reports may be weak, or unreliable. An experiment may have contained confounding variables, for example, or been conducted with a small sample. In practice, one of the factors that determines the high reliability of a respected academic journal is that it reports highly reliable evidence. But nevertheless there is a distinction between the reliability of some evidence, and the reliability of the source *reporting* that evidence.

In fact this distinction is one that has played a crucial role in developing a recent Bayesian approach to epistemology. Bovens and Hartmann (2003) note that when deciding how convinced to be by a particular hypothesis we must make some assessment of both the facts (i.e. whether some evidence is more or less compelling) and the report of those facts (i.e. the source providing the evidence). While this distinction is not one that I have made in this thesis, future work could more carefully consider the relationship between evidential reliability and source reliability. For example, one could ask whether people are preferentially sensitive to one of the types of reliability. In the context of arguments about climate change, are people more attuned to the reliability of climate models, or to the reliability of the scientific institutions that report them?

6.3.5   The Case for Rational Debate

My final suggestion for future research is a more general one, but one that nonetheless falls directly out of the work presented in this thesis. In all of the experiments I have reported, there is a remarkable degree of consistency between people's evaluations of arguments and Bayesian predictions about how strong they *should* be. That is, on the evidence presented here, people are perfectly good at evaluating a range of arguments in a rational way. In addition to providing sound normative guidance for informal argumentation, therefore, Bayesian theory seems also to provide a pretty good descriptive theory of argument evaluation.

However, rationality has something of an image problem. The 'heuristics and biases' literature associated with the work of Kahneman and Tversky (see, e.g., Kahneman, Slovic & Tversky, 1982) seemed to show that people's behaviour in experiments was often not particularly rational, leading to strong conclusions such as:

"In his evaluation of evidence, man is apparently not a conservative Bayesian; he is not Bayesian at all" (Kahneman & Tversky, 1972, p450).

A lot of work has since questioned the conclusions of the heuristics and biases literature from a number of different perspectives (see, e.g., Gigerenzer & Goldstein, 1996; Hilton, 1995: Noveck & Sperber, 2004). Perhaps the fairest assessment would be to say that the extent to which people are Bayesian in their everyday reasoning is open to debate. But the heuristics and biases literature had a powerful impact on perceptions of rationality in the world outside of academia: Anyone vaguely familiar

with the work of Kahneman and Tversky is likely to have absorbed the message that 'man' is *not* a rational animal.

Similarly, a dominant theme in the vast literature on persuasion and communication has been that there are 'two routes' to persuasion – one central, that involves attentive, concentrated processing of the arguments, and one peripheral, which involves a superficial judgement of the most obvious features of the persuasive message (e.g. whether or not a message giver is an expert, regardless of what they say). The popularity of dual-process theories in social psychology (e.g., Eagly & Chaiken, 1980; Petty & Cacioppo, 1984) has meant that much research has focused on ways in which the peripheral route to persuasion can be most effectively utilised. Correspondingly, entire journals dedicated to 'consumer research' have developed, where (ostensibly irrational) factors such as celebrity endorsements, or pleasant music are studied for their persuasive effectiveness (Albarracin & Kumkale, 2003) or effect on purchasing decisions (MacInnas & Park, 1991).

The combination of the 'fact' that people are irrational (c.f. Kahneman, Slovic & Tversky, 1982), with the prevailing belief that it is possible to convince people to make consumer decisions based on little more than a pleasant odour or charming message font, has not just dominated the advertising industry, however. A popular complaint about contemporary political parties in the UK is their apparent preference for style over substance – often referred to as 'political spin'. This style of politics is characterised by a conspicuous lack of rational debate – that is, a lack of argument content to evaluate. Similarly, organisations looking to run effective environmental campaigns are increasingly turning to 'social marketers' for advice on how to promote

their green credentials. For example, a consultancy called Futerra offer 'rules of the game' that claim to give companies the most effective ways of 'greenwashing' their operation (including the pronouncement that 'there is no rational man' – advice they gave to the UK government in a consultation regarding their climate change communications strategy – see, Futerra Sustainability Communications, 2007). Presumably, however, these rules are not nearly as effective as *actually making changes* to the environmental policy of an organisation.

The problem with assuming that people are not rational, that they can be 'tricked' into believing (or buying) all sorts of things, and that the best way of winning an election or running a business is by avoiding engaging in rational debate altogether is that this assumption is the beginning of a vicious circle. Because people are supposedly not rational, and have been shown to respond to non-rational features of persuasive messages, then non-rational means are seen as the best way with which to persuade people. The rational ideal of 'economic man' may have been exposed as a myth, but his replacement seems little better – while assumptions about the average citizen's competence for computing complex statistical inferences may have been misplaced, there are clearly dangers associated with the other extreme.

To illustrate why rational debate cannot be abandoned altogether, consider a government campaign to increase uptake of council tax payment. Having diligently consulted the relevant psychological literature, the junior researcher tasked with putting together an effective advertising strategy concludes that engaging in rational debate with the public is pointless, since psychological experiments have often shown people to be irrational and inattentive to crucial features of persuasive messages.

However, the government would seem to have a fundamental responsibility to *try* to engage their citizens in rational debate – if they do not, then the absence of rational debate is assured. Rational debate has formed a fundamental part of most societies since Ancient Greece. Argumentation is perhaps the primary method by which knowledge is advanced. In politics, law and academia, the construction and evaluation of arguments plays a vital role. But despite its privileged epistemic status, rational argumentation seems to have been abandoned as a tactic for communicating with 'ordinary people'.

The work presented in this thesis, however, suggests that 'ordinary people' are perfectly good at rationally and consistently evaluating a range of different types of arguments on a broad range of topics. While the extent to which undergraduate students are representative of the general population can be questioned, it is certainly not the case that the participants in any of the experiments in this thesis were experts on the topics of the arguments they were evaluating. This suggests that although there may be any number of reasons why, in their everyday lives, people will be motivated to process evidence in a particular way, or be distracted from the rational evaluation of a message, people are *capable* of evaluating arguments and evidence in a rational way.

Given that the dominance of 'spin' in political debate and 'greenwashing' in corporate advertising is generally perceived negatively by the general public, one might wonder whether starting from the assumption that people are – or at least can be – rational, might ultimately be preferable to the prevailing consensus.

6.4 Chapter Summary

The purpose of this chapter has been to summarise the work presented in this thesis, draw some overall conclusions, identify some outstanding questions and make some suggestions for future research. The most important message to take from the work in this thesis is that the Bayesian framework permits a broad range of questions about argument strength to be studied using a very minimal set of theoretical assumptions – specifically, that people's beliefs can be described probabilistically, and that it is rational for people to calibrate their degrees of belief to the probability calculus. Using these assumptions, and drawing on a diverse range of theoretical and empirical work, I have been able to pose novel questions about how people evaluate scientific arguments and evidence, and provide some empirical evidence that bears on the longstanding philosophical dispute over SSAs.

While there are undoubtedly areas where outstanding questions remain – outlined in the sections above – the work in this thesis adds to the growing body of literature (Hahn, Corner & Oaksford, 2006; Hahn & Oaksford, 2006a, 2006b, 2007a, 2007b; Hahn, Oaksford & Corner, 2005; Oaksford & Hahn, 2004) suggesting that the Bayesian approach, as a framework for studying informal argumentation, offers a valuable metric for predicting and measuring argument strength.

References

Adams, S. (2001). Studies of how students and scientists evaluate scientific claims from the world wide web: A method for formulating goals for scientific literacy and critical information literacy. *Unpublished Manuscript.*

Albarracin, D. & Kumkale, T. (2003). Affect as Information in Persuasion: A Model of Affect Identification and Discounting. *Journal of Persoanlity & Social Psychology 84 (3)* 453-469.

Albarracin, D. & Wyer, R.S. (2001). Elaborative and Nonelaborative Processing of a Behaviour-Related Communication. *Personality & Social Psychology Bulletin 27,* 691-705.

Alexy, R. (1989). *A Theory of Legal Argumentation.* Oxford: Clarendon Press.

Allingham, M. (2002). *Choice Theory: A very short introduction.* Oxford: Oxford University Press.

American Family Research Council. (2004). *The Slippery Slope of Same-Sex Marriage.* Washington: Family Research Council.

Anderson, J. R. (1990). *The Adaptive Character of Thought.* Hillsdale, NJ: Lawrence Erlbaum Press.

Anderson, J.R. & Schooler, L.J. (1991). Reflections of the Environment in Memory. *Psychological Science 2 (6)* 396-408.

Areni, C.S. (2002). The Proposition-Probability Model of Argument Strength Structure and Message Acceptance. *Journal of Consumer Research 29*, 168-187.

Areni, C.S. & Lutz, R.J. (1988). The Role of Argument Quality in the Elaboration Likelihood Model. *Advances in Consumer Research 15*, 197-203.

Aristotle (350BC). On Sophistical Refutations: On coming-to-be and passing away. On the cosmos. In *Aristotle*: Translated by E.S. Forster & D.J. Farley (1955). Cambridge, MA: Harvard University Press.

Armendt, B. (1993). Dutch Books, Additivity and Utility Theory. *Philosophical Topics 21 (1)* 1-20.

Audi, R. (2002). Epistemology: A Contemporary Introduction. London: Routledge.

Bailenson, J. & Rips, L. J. (1996). Informal reasoning and burden of proof. *Applied Cognitive Psychology, 10,* 3-16.

Bacchus, F., Kyburg, H.E. & Thalos, M. (1990). Against Conditionalization. *Synthese 85 (3)* 475-506.

Bickerstaff, K., Lorenzoni, I., Pidgeon, N.F., Poortinga, W. and Simmons, P. (in press) Re-framing nuclear power in the UK energy debate: nuclear power, climate change mitigation and radioactive waste, *Public Understanding of Science.*

Biro, J. & Siegel, H. (2006). In Defense of the Objective Epistemic Approach to Argumentation. *Informal Logic 26 (1)* 91-101.

Bishop, M.A. & Trout, J.D. (2005). *Epistemology and the Psychology of Human Judgment.* New York: Oxford University Press.

Boger, G. (2005). Subordinating truth – is *acceptability* acceptable? *Argumentation, 19,* 187-238.

Bonnefon, J.F. & Hilton, D.J. (2004). Consequential Conditionals: Invited and Suppressed Inferences From Valued Outcomes. *Journal of Experimental Psychology; Learning, memory & Cognition 30 (1)* 28-39.

Bovens, L. & Hartmann, S. (2003). *Bayesian Epistemology.* Oxford: Oxford University Press.

Brante, T., Fuller, S. & Lynch, W. (Eds) (1993). *Controversial Science: From Content to Contention.* New York: New York Press.

Chaiken, S. (1980). Heuristic Versus Systematic Information Processing and the Use of Source Versus Message Cues in Persuasion. *Journal of Personality and Social Psychology 39 (5)* 752-766.

Chater, N. & Oaksford, M. (2000). The Rational Analysis of Mind and Behaviour. *Synthese 122*, 93-131.

Chater, N., & Oaksford, M. (Eds.) (2008). *The probabilistic mind: Prospects for Bayesian cognitive science*. Oxford: Oxford University Press.

Christenesen, D. (1996). Dutch-Book Arguments Depragmatized: Epistemic Consistency for Partial Believers. *The Journal of Philosophy 93 (9)* 450-479.

Cohen, L.J. (1981). Can Human Irrationality be Experimentally Demonstrated? *Behavioural and Brain Sciences 4,* 317-370.

Collins, H.M. & Evans, R. (2007). *Rethinking Expertise*. Chicago: University of Chicago Press.

Collins, H. M., & Pinch, T. J., (1993) *The Golem: What You Should Know About Science*. Cambridge & New York: Cambridge University Press.

Cooper, J. & Fazio, R. H. (1984). A new look at dissonance theory. In L. Berkowitz (Ed). *Advances in Experimental Social Psychology 17*. New York: Academic Press.

Corner, A. & Hahn, U. (2007). Evaluating the Meta-Slope: Is there a Slippery Slope Argument against Slippery Slope Arguments? *Argumentation 21 (4)* 349-359.

Corner, A., Hahn, U. & Oaksford, M. (2006). The Slippery Slope Argument: Probability, Utility and Category Boundary Re-appraisal. *Proceedings of The 28th Annual Conference of the Cognitive Science Society,* 1145-1151. Vancouver.

Davidson, B. & Pargetter, R. (1985). In Defence of the Dutch Book Argument. *Canadian Journal of Philosophy 15 (3)* 405-424.

de Finetti, B. (1974). *Theory of Probability.* New York: Wiley.

de Vries, N., Ruiter, R. & Leegwater, Y. (2002). Fear Appeals in persuasive communication. In: G. Bartels & W. Nelissen. (Eds). *Marketing for Sustainability: Towards Transactional Policy Making.* Amsterdam: IOS Press.

Dickerson, C.A., Thibodeau, R., Aronson, E. & Miller, D. (1992). Using Cognitive Dissonance to Encourage Water Conservation. *Journal of Applied Social Psychology 22 (11)* 841-854.

Dillard, J.P., Hunter, J.E. & Burgoon, M. (1984). Sequential Request Persuasion Strategies: Meta-Analysis of Foot-In-The-Door and Door-In-The-Face. *Human Communication Research 10 (4)* 461-488.

Driver, R., Newton, P. & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education 84,* 287-312.

Eagly, A.H. & Chaiken, S. (1993). *The psychology of attitudes.* Orlando, FL: Harcourt Brace.

Edwards, W. (1961). Behavioural Decision Theory. *Annual Review of Psychology 12,* 473-498.

Edwards, W. & Tversky, A. (Eds). (1967). *Decision Making.* Middlesex: Penguin.

Elqayam, S. (2003). Norm, error and the structure of rationality: The case study of the knight-knave paradigm. *Semiotica 147,* 265-289.

Elqayam, S. (2007). Normative Rationality and the Is-ought Fallacy. *Proceedings of the 2ⁿᵈ Meeting of the European Cognitive Science Society,* 294-299.

Enoch, D. (2001). Once you start using Slippery Slope Arguments, you're on a very Slippery Slope. *Oxford Journal of Legal Studies 21 (4)* 629-647.

Erduan, S., Simon, S. & Osborne, J. (2004). TAPping into Argumentation: Developments in the Application of Toulmin's Argument Pattern for Studying Science Discourse. *Science Education 88,* 915-933.

Evans, J. St. B.T (2002). Logic and human reasoning: an assessment of the deduction paradigm. *Psychological Bulletin 128(6)* 978-996.

Evans, J.St.B.T & Over, D.E. (1996). *Rationality and Reasoning*. Hove, UK:

Psychology Press.

Evans, J.St.B.T. & Over, D.E. (2004). *If.* Oxford: Oxford University Press.

Evans, J.St.B.T., Over, D.E. & Handley, S.J. (2005). Suppositions, extensionality and

conditionals: A critique of the mental model theory of Johnson-Laird & Byrne (2002).

Psychological Review 112, 1040-1052.

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford: Stanford

University Press.

Finnis, J. (1980). *Natural Law & Natural Right*. Oxford: Oxford University Press.

Fishbein, M. & Ajzen, I. (1975). *Belief, Attitude, Intention & Behaviour*. California:

Addison – Wesley Publishing Company.

Fishbein, M. & Ajzen, A. (1981). Acceptance, Yielding and Impact: Cognitive

Processes in Persuasion. In: Petty, R.E, Ostrom, T.M, Brock, T.C. (Eds) (1981).

Cognitive Responses in Persuasion. New Jersey: Lawrence Erlbaum.

Freedman, J.L. & Fraser, S.C. (1966). Compliance without pressure: The foot-in-the-

door technique. *Journal of Personality & Social Psychology 4 (2)* 195-202.

Friedman, S.M., Dunwoody, S. & Rogers, C.L. (Eds). (1999). *Communicating Uncertainty: Media Coverage of New and Controversial Science*. New Jersey: Lawrence Erlbaum.

Fugelsang, J.A., Stein, C.B., Green, A.E. & Dunbar, K.N. (2004). Theory and Data Interactions of the Scientific Mind: Evidence From the Molecular and the Cognitive Laboratory. *Canadian Journal of Experimental Psychology 58 (2)* 86-95.

Futerra Sustainability Communications (2007). Rules of the Game: Evidence Base for the Climate Change Strategy. http://www.futerra.co.uk/downloads/rulesofthegame.pdf

Gamut, L.T.F. (1991). *Logic, Language and Meaning Volume 1*. Chicago: University of Chicago Press.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: beyond 'heuristics and biases'. *European Review of Social Psychology 2 (1)* 83-115.

Gigerenzer, G. & Edwards, A. (2003). Simple tools for understanding risks: from innumeracy to insight. *British Medical Journal 327 (7417)* 741–744.

Gigerenzer, G. & Goldstein, D.G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review 104*, 650-669.

Gigerenzer, G. & Selten, R.S. [Eds] (2002). *Bounded Rationality: The Adaptive Toolbox*. Massachusetts: MIT.

Gigerenzer, G. & Todd, P.M. (1999). *Simple heuristics that make us smart*. Oxford: Oxford University Press.

Gilbert, M.A. (2007). Natural Normativity: Argumentation Theory as an Engaged Discipline. *Informal Logic 27 (2)* 149-161.

Gilovich, T., Griffin, D.W. & Kahneman, D. (2002). Heuristics & Biases: The psychology of intuitive judgment. Cambridge: Cambridge University Press.

Goldman, A.I. (2003). An epistemological approach to argumentation. *Informal Logic, 23*, 51-63.

Govier, T. (1982). What's wrong with slippery slope arguments. *Canadian Journal of Philosophy 12*, 303-316.

Govier, T. (1987) The Fallacy Behind Fallacies – Reply to Massey. In H.V. Hansen. & R.C. Pinto. (1995). *Fallacies: Classical and Contemporary Readings*. Pennsylvania: Pennsylvania State University Press.

Green, D. M. & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Gregory, J. & Miller, S. (1998). *Science in Public: Communication, Culture & Credibility*. Cambridge: Basic Books.

Grice, H.P. (1975). Logic and Conversation. In D. Davidson & G. Harman (Eds). *The Logic of Grammar*. Encino, California: Dickenson.

Haddock, G. & Maio, G.R. (Eds) (2004) *Contemporary Perspectives on the Psychology of Attitudes*. NY: Psychology Press.

Hahn, U. & Oaksford, M. (2006a). A Bayesian Approach to Informal Fallacies. *Synthese 152 (2)* 207-237.

Hahn, U. & Oaksford, M. (2006b) Why a normative theory of argument strengthand why might one want it to be Bayesian? *Informal Logic, 26*,1-24.

Hahn, U. & Oaksford, M. (2007a). The Rationality of Informal Argumentation: A Bayesian Approach to Reasoning Fallacies. *Psychological Review 114 (3)* 704-732.

Hahn, U. & Oaksford, M. (2007b). The burden of proof and its role in argumentation. *Argumentation 21*, 39-61.

Hahn, U, Oaksford, M. & Corner, A. (2005) Circular arguments, begging the question and the formalization of argument strength. *Proceedings of AMKCL05 –Adaptive Knowledge Representation and Reasoning*. Helsinki.

Hajek, A. (2005). Scotching Dutch Books? *Philosophical Perspectives, 19,* 139-151.

Hamblin, C.L. (1970). *Fallacies*. London: Methuen.

Hart, H.L.A. (1961). *The Concept of Law*. Oxford: Oxford University Press.

Heysse, T. (1997). Why logic doesn't matter in the (philosophical) study of argumentation. *Argumentation, 11*, 211-224.

Hilton, D.J. (1995). The Social Context of Reasoning: Conversational Inference and Rational Judgment. *Psychological Bulletin 118 (2)* 248-271.

Hoeken, H. (2001a). Convincing Citizens. The Role of Argument Quality. In D. Janssen & R. Neutelings (Eds). *Reading and Writing Public Documents*. Amsterdam: Benjamins.

Hoeken, H. (2001b). Anecdotal, Statistical and Causal Evidence: Their Perceived and Actual Persuasiveness. *Argumentation 15,* 425-437.

Hookway, C. (1993). Epistemic Norms and theoretical Deliberation. In J. Dancy. (Ed). (2000). *Normativity*. Oxford: Blackwell.

Hoog, N., Stroebe, W. & de Wit, J.B.F. (2005). The impact of Fear Appeals on processing and acceptance of action recommendations. *Personality & Social Psychology Bulletin, 31 (1)* 24-33.

Holtug, N. (1993). Human Gene Therapy: Down the Slippery Slope. *Bioethics 7*, 402-

419.

Howson, C. & Urbach, P. (1996). *Scientific Reasoning: The Bayesian Approach.*

Chicago: Open Court.

Hume, D. (1740). *A Treatise of Human Nature* (1967 edition). Oxford: Oxford

University Press.

Intergovernmental Panel on Climate Change (2007). Climate Change 2007: Synthesis

Report.

Irwin, A. & Wynne, B. (Eds.) (1996). *Misunderstanding Science? The public*

reconstruction of science and technology. Cambridge: Cambridge University Press.

Jimenez-Alexandre, M.P. (2002). Knowledge producers or knowledge consumers?

Argumentation and decision making about environmental management. *International*

Journal of Science Education 24 (11) 1171-1190.

Jimenez-Alexandre, M.P., Rodriguez, A.B. & Duschl, R.A. (2000). "Doing the

lesson" or "doing science:" Argument in High School genetics. *Science Education 84*,

757-792.

Johnson, B. T., Maio, G. R., & Smith-McLallen, A. (2005). Communication and

attitude change: Causes, processes, and effects. In D. Albarracin, B. T. Johnson, & M.

P. Zanna (Eds.), *Handbook of attitudes and attitude change* (617–669). Mahwah, NJ: Erlbaum.

Johnson, B.T., Smith-McLallen, A., Killeya, L.A. & Levin, K.D. (2004). Truth or Consequences: Overcoming Resistance to Persuasion with Positive Thinking. In E.S. Knowles & J.A. Linn (Eds) (2004). *Resistance and Persuasion*. New Jersey: Laurence Erlbaum.

Johnson, R.H. (2000). *Manifest rationality: a pragmatic theory of argument.* Mahwah, NJ: Hillsdale.

Johnson-Laird, P. N. (1983). *Mental Models.* Cambridge: Cambridge University Press.

Johnson-Laird, P.N. & Bara, B.G. (1984). Syllogistic Inference. *Cognition 16 (1)* 1-61.

Johnson-Laird, P.N. & Byrne, R.M.J. (2002). Conditionals: a theory of meaning, pragmatics and inference. *Psychological Review 109*, 646-678.

Kahneman, D., Slovic, P. & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.

Kahneman, D. & Tversky, A. (1972). Subjective Probability: A judgement of representativeness. *Cognitive Psychology 3,* 430-454.

Kahneman, D. & Tversky, A. (1979). Prospect Theory: An Analysis of Decision

Making under Risk. *Econometrica 47 (2)* 263-292.

Kantola, S.J., Syme, G.J. & Campbell, N.A. (1984). Cognitive dissonance and energy

conservation. *Journal of Applied Psychology 69*, 416–421.

Kaplan, M. F. (1971). Dispositional effects and weight of information in

impression formation. *Journal of Social Psychology 18*, 279–284.

Keeney, R.L. & Raiffa, H. (1976). *Decisions with multiple objectives: Preferences

and value tradeoffs*. New York: Wiley.

Kelsen, H. (1941). The Pure Theory of Law & Analytical Independence. *Harvard

Law Review 55*, 44-66.

Kennedy, R. & Chihara, C. (1979). The Dutch Book Argument: Its logical flaws, its

subjective sources. *Philosophical Studies 36*, 19-33.

Kirk, R.E. (1995). *Experimental Design – Procedures for the Behavioural Sciences*.

Brooks/Cole: London.

Klaczynski, P. (2000). Motivated scientific reasoning biases, epistemological biases,

and theory polarisation: A two process approach to adolescent cognition. *Child

Development 71*, 1347-1366.

Knowles, J. (2003). *Norms, Naturalism and Epistemology*. New York: Palgrave

Macmillan.

Kolsto, S.D. (2001). "To trust of not to trust..." – pupils' ways of judging information

encountered in a socio-scientific issue. *International Journal of Science Education 23*

(9), 877-901.

Kolsto, S.D., Bungum, B., Arneses, E., Isnes, A., Kristensen, T., Mathiassen, K.,

Mestad, I., Quale, A., Tonning, A.S.V. & Ulvik, M. (2006). Science Students' Critical

Examination of Scientific Information related to Socioscientific Issues. *Science*

Education 90, 632-655.

Korb, K.B. & Nicholson, A.E. (2004). *Bayesian Artificial Intelligence*. Florida: CRC

Press.

Korb, K.B. (2004). Bayesian Informal Logic and Fallacy. *Informal Logic 24,* 41-70.

Korpan, C.A., Bisanz, G.L., Bisanz, J. & Henderson, J.M. (1997). Assessing Literacy

in Science: Evaluation of Scientific News Briefs. *Science Education 81,* 515-532.

Kortland, K. (1996). An STS case study about students' decision making on the waste

issue. *Science Education 80*, 673-689.

Kruglanski, A.W., Fishbach, A., Erb, H. P., Pierro, A., & Mannetti, L. (2004). The

Parametric Unimodel as a Theory of Persuasion. In G. Haddock and G. R. Maio

(Eds.), *Contemporary Perspectives on the Psychology of Attitudes*. NY: Psychology Press.

Kuhn, D., Cheney, R. & Weinstock, M. (2001). The development of epistemological understanding. *Cognitive Development 15,* 309-328.

Kuhn, D., Shaw, V. & Felton, M. (1997). Effects of Dyadic Interaction on Argumentative Reasoning. *Cognition & Instruction 15 (3)* 287-315.

Kuhn, D. & Udell, W. (2003). The Development of Argument Skills. *Child Development 74 (5)* 1245-1260.

Kuhn, T. S. (1970). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

Lagnado, D.A. & Harvey, N. (in press). The impact of discredited evidence. *Psychonomic Bulletin & Review.*

Lakoff, G. (1987). Cognitive Models and Prototype Theory. In, U. Neisser (Ed). *Concepts and conceptual development: Ecological and Intellectual Factors.* Cambridge: Cambridge University Press.

Lamb, D. (1988). *Down the Slippery Slope – Arguing in Applied Ethics.* London: Croom Helm.

Lamberts, K. (1995). Categorization under time pressure. *Journal of Experimental Psychology: General, 124,* 161-180.

Lando, O. & Beale, H. (2000). *Principles of European Contract Law.* Boston: Kluwer Law International.

Launis, V. (2002). Human Gene Therapy & the Slippery Slope Argument. *Medicine, Health Care and Philosophy 5,* 169-179.

Lindley, D.V. (1982). Scoring rules and the inevitability of probability. *International Statistical Review 50,* 1-26.

Lode, E. (1999). Slippery Slope Arguments and legal Reasoning. *California Law Review 87,* 1468-1543.

Lorenzoni, I. and Pidgeon, N.F (2006) Public views on climate change: European and USA perspectives. *Climatic Change, 77,* 73-95

MacInnas, D.J. & Park, C.W. (1991). The differential role of characteristics of music on high- and low-involvement consumers' processing of ads. *Journal of Consumer Research 18,* 161-173.

Maio, G.R. & Haddock, G. (2007). Attitude Change. In A .W. Kruglanski & E. T. Higgins (Eds.), *Social psychology: Handbook of basic principles* (Vol. 2). New York, NY: Guilford Press.

Manktelow, K.I. & Over, D.E. (1991). Social roles and utilities in reasoning with deontic conditionals. *Cognition 39(2)* 85-105.

McGuire, W.J. (1960). Cognitive Consistency and Attitude Change. *Journal of Abnormal and Social Psychology 60 (3)* 345-358.

McGuire, W.J. (1981). The Probabilogical Model of Cognitive Structure and Attitude Change. In: R.E. Petty, T.M. Ostrom, T.C. Brock (Eds). *Cognitive Responses in Persuasion*. New Jersey: Lawrence Erlbaum.

Nelson, J.D. (2005). Finding useful questions: On Bayesian Diagnosticity, Probability, Impact and Information Gain. *Psychological Review 112 (4)* 979-999.

Nelson, J.D. (2008). Towards a rational theory of human information acquisition. In: N. Chater & M. Oaksford. [Eds.]. *The probabilistic mind: prospects for rational models of cognition.* Oxford: Oxford University Press.

Newton, P. (1999). The Place of Argumentation in the Pedagogy of School Science. *International Journal of Science Education 21 (5)* 553-576.

Nickerson, R.S. (2007). *Aspects of Rationality: Reflections on What It Means To Be Rational and Whether We Are*. New Jersey: Psychology Press.

Nisbett, R.E. & Ross, L. (1980). *Human Inference: Strategies and Shortcomings of Social Judgement*. NJ: Prentice Hall.

Nisbett, R.E., Zukier, H. & Lemley, R.E. (1981). The dilution effect: Non-diagnostic information weakens the implications of diagnostic information. *Cognitive Psychology 13*, 248-277.

Norris, S.P., Phillips, L.M. & Korpan, C.A. (2003). University students' interpretation of media reports of science and its relationship to background knowledge, interest and reading difficulty. *Public Understanding of Science 12*, 123-145.

Nosofsky, R.M. (1986) Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39-57.

Nosofsky, R.M. (1988a). Exemplar-based accounts of the relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory and Cognition, 14*, 700-708.

Nosofsky, R.M. (1988b). Similarity, frequency and category representation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 54-65.

Noveck, I.A. & Sperber, D. (Eds). (2004). *Experimental Pragmatics*. New York: Palgrave Macmillan.

O'Keefe, D. J. (1995). Argumentation studies and dual-process models of persuasion. In F. H. van Eemeren, R. Grootendorst, J. A. Blair, & C. A. Willard (Eds.), *Proceedings of the third ISSA conference on argumentation, vol. 1: Perspectives and approaches* (3-17). Amsterdam: Sic Sat.

O'Keefe, D.J. (1997a). Standpoint explicitness and persuasive effect: A meta-analytic review of the effects of varying conclusion articulation in persuasive messages. *Argumentation & Advocacy 34*, 1-12.

O'Keefe, D.J. (1997b). Justification explicitness and persuasive effect: A meta-analytic review of the effects of varying support articulation in persuasive messages. *Argumentation & Advocacy 35*, 61-75.

O'Keefe, D.J. (2003). The Potential Conflict Between Normatively Good Argumentative Practice and Persuasive Success. In F.H. van Eemeren, J. Anthony Blair, C.A. Willard & A. Francisca Snoeck Henkemans (Eds). *Anyone Who Has A View: Theoretical Contributions to the Study of Argumentation.* Dordrecht: Kluwer Academic Publishers.

O'Keefe, D. J. (2005). News for argumentation from persuasion effects research: Two cheers for reasoned discourse. In C. A. Willard (Ed.), *Selected papers from the thirteenth NCA/AFA conference on argumentation* (pp. 215-221). Washington, DC: National Communication Association.

O'Keefe, D. J., & Jackson, S. (1995). Argument quality and persuasive effects: A review of current approaches. In S. Jackson (Ed.), *Argumentation and values: Proceedings of the ninth Alta conference on argumentation* (88-92). Annandale, VA: Speech Communication Association.

Oakley, J. & Cocking, D. (2005). Consequentialism, Complacency and Slippery Slope Arguments. *Theoretical Medicine & Bioethics 26*, 227-239.

Oaksford, M. & Chater, N. (1991). Against logicist cognitive science. *Mind & Language 6*, 1-38.

Oaksford, M. & Chater, N. (1998). *Rationality in an uncertain world: essays on the cognitive science of human reasoning.* Sussex: Psychology Press.

Oaksford, M. & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences 5 (8)*, 349-357.

Oaksford, M. & Chater, N. (2003). Conditional Probability and the Cognitive Science of Conditional Reasoning. *Mind & Language 18 (4)*, 359–379.

Oaksford, M. & Chater, N. (2007). *Bayesian Rationality: The probabilistic approach to human reasoning.* Oxford: Oxford University Press.

Oaksford, M. & Hahn, U. (2004). A Bayesian approach to the argument from ignorance. *Canadian Journal of Experimental Psychology, 58,* 75-85.

Osherson, D., Smith, E.E., Wilkie, O., Lòpez, A. & Shafir, E. (1990). Category based induction. *Psychological Review 97 (2)* 185-200.

Over, D.E., Manktelow, K.I. & Hadjichristidis, C. (2004). Conditions for the acceptance of deontic conditionals. *Canadian Journal of Experimental Psychology 58 (2)* 96-105.

Oxford English Dictionary. (2006). Oxford: Oxford University Press.

Parsons, S. (2001). *Qualitative methods for reasoning under uncertainty*. Massachusetts: MIT Press.

Patronis, T., Potari, D. & Spiliotopolou, V. (1999). Students' argumentation in decision-making on a socio-scientific issue: implications for teaching. *International Journal of Science Education 21 (7)* 745-754.

Patt, A. (2007). Assessing model-based and conflict-based uncertainty. *Global Environmental Change 17*, 37-46.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufman.

Perelman, C., & Olbrechts-Tyteca, L. (1969). *The new rhetoric: A treatise on argumentation*. Notre Dame, IN: University of Notre Dame Press.

Petty, R.E. & Cacioppo, J.T. (1979). Issue Involvement Can Increase or Decrease Persuasion by Enhancing Message-Relevant Cognitive Responses. *Journal of Personality and Social Psychology 37*, 1915-1926

Petty, R.E. & Cacioppo, J.T. (1984). The Effects of Involvement on responses to

Argument Quantity and Quality: Central and Peripheral Routes to Persuasion. *Journal

of personality and Social psychology 46 (1)* 69-81.

Petty, R.E. & Cacioppo, J.T. (1986). *Communication and Persuasion: Central and

Peripheral Routes to Attitude Change.* New York: Springer-Verlag.

Petty, R. E., Cacioppo, J. T., & Goldman, R. (1981). Personal involvement

as a determinant of argument-based persuasion. *Journal of Personality

and Social Psychology, 41,* 847–855.

Petty, R.E. & Wegener, D. T. (1991). Thought systems, argument quality, and

persuasion. Wyer, R. S., & Srull, T. K. (Eds.) (1991). Advances in Social Cognition

(Vol. 4, pp. 147-161). Hillsdale, NJ: Erlbaum.

Phillips, L.M. & Norris, S.P. (1999). Interpreting popular reports of science: what

happens when the reader's world meets the world on paper? *International Journal of

Science Education 21 (3)* 317-327.

Pidgeon, N., Kasperson, R.E. and Slovic, P. (2003) *The Social Amplification of Risk.*

Cambridge University Press, Cambridge.

Pidgeon, N.F. & Rogers-Hayden, T. (2007) Opening up nanotechnology dialogue

with the publics: risk communication or 'upstream engagement'? *Health, Risk and

Society,* 9, 191-210.

Pollack, H.N. (2005). *Uncertain Science...Uncertain World*. Cambridge: Cambridge University Press.

Popper, K.R. (1959). *The logic of scientific discovery*. London: Hutchison.

Poortinga, W., & Pidgeon, N. F. (2003). *Public perceptions of risk, science and governance*. Norwich: UEA/MORI.

Pothos, E. & Hahn, U. (2000). So concepts aren't definitions, but do they have necessary or sufficient features? *British Journal of Psychology, 91*, 439-450.

Pratt. J, Raiffa.H, & Schlaifer, R. (1995). *Introduction to Statistical Decision Theory*. Massachusetts: MIT Press.

Prislin, R. & Roullette, J. (1996). When it is Embedded, it is Potent: Effects of General Attitude Embeddedness on Formation of Specific Attitudes and Behavioural Intentions. *Personality & Social Psychology Bulletin 22 (8)* 845-861.

Quine, W.V.O. (1969) *Ontological Relativity and Other Essays*. New York: Columbia University Press.

Railton, P. (2000). Normative Force and Normative Freedom: Hume & Kant. In J. Dancy (Ed). (2000). *Normativity*. Oxford: Blackwell

Ramsey, F.P. (1931). *The Foundations of Mathematics and other Logical Essays*. London: Kegan, Paul, Trench, Trubner & Co.

Ratcliffe, M. (1999). Evaluation of abilities in interpreting media reports of scientific research. *International Journal of Science Education 21 (10)* 1085-1099.

Ricco, R.B. (2007). Individual differences in the analysis of informal reasoning fallacies. *Contemporary Educational Psychology 32*, 459-484.

Rips, L.J. (1998). Reasoning and Conversation. *Psychological Review 105*, 411-441.

Rips, L.J. (2001). Two kinds of reasoning. *Psychological Science 12*, 129-134.

Rizzo, M.J. & Whitman, D.G. (2003). The Camel's Nose Is In The Tent: Rules, Theories and Slippery Slopes. *UCLA Law Review 51*, 539 – 592.

Rosenthal, R. (1984). *Applied Social Science Research Methods Series: Meta-analytic procedures for social research.* London: Sage Publications.

Rowbottom, D.P. (2007). The Insufficiency of the Dutch Book Argument. *Studia Logica 87*, 65-71.

Sadler, T.D. (2004). Informal Reasoning Regarding Socioscientific Issues: A Critical Review of Research. *Journal of Research in Science Teaching 41 (5)* 513-536.

Savage, L. J. (1954). *The foundations of statistics.* New York: John Wiley.

Schum, D.A. (1994) *Evidential Foundations of Probabilistic Reasoning*. New York: John Wiley.

Schwarz, N. (1996). *Cognition & Communication: Judgemental Biases, Research Methods & The Logic of Conversation*. Hillsdale, NJ: Erlbaum.

Schwarz, N. & Bless, H. (1992). Constructing reality and its alternatives: Assimilation and contrast effects in social judgment. In L.L. Martin & A. Tesser (Eds). *The construction of social judgment* (217-245). Hillsdale, NJ: Erlbaum.

Segnit, N. & Ereaut, G. (2007). *Warm Words II: How the climate story is evolving and the lessons we can learn for encouraging public action*. London: Institute for Public Policy Research.

Sherif, C.W., Sherif, M. & Nebergall, R.E. (1965). *Attitude and Attitude Change: The Social Judgment-Involvement Approach*. Philadelphia: W.B. Saunders.

Sibler, D.S. (1999). Dutch Books and Agent Rationality. *Theory and Decision 47*, 247-266.

Siegel, H. & Biro, J. (1997). Epistemic Normativity, Argumentation & Fallacies. *Argumentation 11*, 277-292.

Siegel, H. & Biro, J. (2008). Rationality, Reasonableness, and Critical Rationalism: Problems with the Pragma-dialectical View. *Argumentation 22*, 191-203.

Simon, H.A. (1982). *Models of Bounded Rationality*, Vols. 1, 2. Cambridge, MA: MIT Press.

Simon, S., Erduran, S. & Osborne, J. (2002). Enhancing the Quality of Argumentation in School Science. *Proceedings of the Annual Meeting of the National Association for Research in Science Teaching: New Orleans, USA.*

Slob, W.H. (2002) How to distinguish good and bad arguments: dialogico-rhetorical normativity. *Argumentation, 16,* 179-196.

Sloman, S.A. (1993). Feature-based induction. *Cognitive Psychology 25,* 231–280.

Slovic, P. & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgement. *Organizational Behavior & Human Processes 6,* 649-744.

Slovic, P. & Tversky, A. (1974). Who Accepts Savage's Axiom? *Behavioral Science 19 (6)* 368-373.

Stanovich, K.E. (1999). *Who Is Rational? Studies of Individual Differences in Reasoning.* New Jersey: Lawrence Erlbaum.

Stanovich, K.E. & West, R.F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Science 23,* 645-726.

Stapel, D. A., & Winkielman, P. (1998). Assimilation and contrast as a function of context target similarity, distinctness, and dimensional relevance. *Personality and Social Psychology Bulletin, 24*, 634-646.

Stewart, N., Chater, N., Stott, H.P. & Reimers, S. (2003). Prospect Relativity: How Choice Options Influence Decision Under Risk. *Journal of Experimental Psychology: General 132 (1)* 23-46.

Stich, S.P. (1985). Could man be an irrational animal? *Synthese 64,* 115-135.

Stich, S.P. (1990). *The Fragmentation of Reason.* Cambridge, MA: MIT Press.

Takao, A.Y. & Kelly, G.J. (2003). Assessment of Evidence in Unversity Students' Scientific Writing. *Science & Education 12,* 341-363.

Tentori, K., Crupi, V., Bonini, N. & Osherson, D. (2007). Comparison of confirmation measures. *Cognition 103,* 107-119.

Thompson, V.A, Evans J. St. B.T. & Handley, S.J. (2005). Persuading and dissuading by conditional argument. *Journal of Memory & Language 53,* 238-257.

Toulmin, S. (1958). *The Uses of Argument.* Cambridge: Cambridge University Press.

Tversky, A. and Kahneman, D. (1982). Judgments of and by representativeness. In D. Kahneman, P. Slovic & A. Tversky (Eds). *Judgment under uncertainty: Heuristics and biases.* Cambridge, UK: Cambridge University Press.

Tversky, A. & Kahneman, D. (1983). Extensional versus Intuitive Reasoning: The

Conjunction Fallacy in Probability Judgment. *Psychological Review 90 (4)* 293-215.

van Der Burg, W. (1991). The Slippery Slope Argument. *Ethics 102,* 42-45.

van Eemeren, F.H. & Grootendorst, R. (2004). *A systematic theory of argumentation
– the pragma-dialectical approach.* Cambridge: Cambridge University Press.

van Enschot-Van Dijk, R., Hustinx, L. & Hoeken, H. (2003). The Concept of

Argument Quality in the Elaboration Likelihood Model. In F.H. van Eemeren, J.

Anthony Blair, C.A. Willard & A. Francisca Snoeck Henkemans (Eds). *Anyone Who

Has A View: Theoretical Contributions to the Study of Argumentation.*(319-333).
Dordrecht: Kluwer Academic Publishers.

Volokh, E. (2003). The Mechanisms of The Slippery Slope. *Harvard Law Review
116,* 1026-1137.

Volokh, E. & Newman, D. (2003). In Defense of the Slippery Slope. *Legal Affairs
March/April,* 21-23.

von Aufschaiter, C., Erduran, S., Osborne, J. & Simon, S. (2008) Arguing to Learn

and Learning to Argue: Case Studies of How Students' Argumentation Relates to

Their Scientific Knowledge. *Journal of Research in Science Teaching 45 (1)* 101-131.

von Neumann, J. & Morgenstern, O. (1944). *Theory of Games and Economic Behaviour*. Princeton: Princeton University Press.

Waidacher, C. (1997). Hidden Assumptions in the Dutch Book Argument. *Theory and Decision 43*, 293-312.

Walton, D. (1989). *Informal Logic: A Handbook for Critical Argumentation*. Cambridge: Cambridge University Press.

Walton, D. (1992a). Non-fallacious Arguments From Ignorance. *American Philosophical Quarterly 29 (4)* 381-387.

Walton, D. (1992b). *Slippery Slope Arguments*. Oxford: Clarendon Press.

Walton, D. (1995). *A Pragmatic Theory of Fallacy*. University of Alabama Press: Alabama.

Walton, D. (1998). *The New Dialectic – Conversational Contexts of Argument*. Toronto: University of Toronto Press.

Walton, D.N. (2005). Begging the Question in Arguments Based on Testimony. *Argumentation 19*, 85-113.

Walton, D. (2006). Epistemic and Dialectical Models of Begging the Question. *Synthese 152*, 237–284.

Wason, P.C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology 20, 273-281.*

Wegener, D.T. & Petty, R.F. (1995). Flexible correction processes in social judgment: The role of naïve theories in corrections for perceived bias. *Journal of Personality & Social Psychology 68 (1)* 36-51.

Woods, J. & Walton, D. (1982). *Argument: The Logic of the Fallacies.* Toronto: McGraw-Hill Ryerson.

Woods, J, Irvine, A. & Walton, D. (2004). *Critical Thinking, Logic & The Fallacies.* Toronto: Prentice Hall.

Wyer Jr, R.S. & Goldberg, L. (1970). A Probabilistic Analysis of the Relationships among Beliefs and Attitudes. *Psychological Review 77 (2)* 100-120.

Zehr, S. (2000). Public representations of scientific uncertainty about global climate change. *Public Understanding of Science 9*, 85-103.