Philosophies of Probability

Jon Williamson

Draft of July 21, 2008.
In A. Irvine (ed.): *Handbook of the Philosophy of Mathematics*,
Volume 4 of the Handbook of the Philosophy of Science, Elsevier.

Abstract

This chapter presents an overview of the major interpretations of probability followed by an outline of the objective Bayesian interpretation and a discussion of the key challenges it faces. I discuss the ramifications of interpretations of probability and objective Bayesianism for the philosophy of mathematics in general.

Contents

§1
Introduction

The concept of probability motivates two key questions.

First, how is probability to be defined? Probability was axiomatised in the first half of the 20th century (Kolmogorov, 1933); this axiomatisation has by now become well entrenched, and in fact the main leeway these days is with regard to the type of domain on which probability functions are defined. Part I introduces three types of domain: variables (§2), events (§3), and sentences (§4).

Second, how is probability to be applied? In order to know how probability can be applied we need to know what probability means: how probabilities can be measured and how probabilistic predictions say something about the world. Part II discusses the predominant interpretations of probability: the frequency (§6), propensity (§7), chance (§§8, 10), and Bayesian interpretations (§9).

In Part III, we shall focus on one interpretation of probability, objective Bayesianism, and look more closely at some of the challenges that this interpretation faces. Finally, Part IV draws some lessons for the philosophy of mathematics in general.

Part I
Frameworks for Probability

§2
Variables

The most basic framework for probability involves defining a probability function relative to a finite set V of variables, each of which takes finitely many possible values. I shall write $v@V$ to indicate that v is an assignment of values to V.

A *probability function* on V is a function P that maps each assignment $v@V$ to a non-negative real number and which satisfies *additivity*:

$$\sum_{v@V} P(v) = 1.$$

This restriction forces each probability $P(v)$ to lie in the unit interval $[0, 1]$.

The *marginal probability function* on $U \subseteq V$ induced by probability function P on V is a probability function Q on U which satisfies

$$Q(u) = \sum_{v@V, v \sim u} P(v)$$

for each $u@U$, and where $v \sim u$ means that v is consistent with u, i.e., u and v assign the same values to $U \cap V = U$. The marginal probability function Q on U is uniquely determined by P. Marginal probability functions are usually thought of as extensions of P and denoted by the same letter P. Thus P can be construed as a function that maps each $u@U \subseteq V$ to a non-negative real number. P can be further extended to assign numbers to conjunctions tu of assignments where $t@T \subseteq V, u@U \subseteq V$: if $t \sim u$ then tu is an assignment to $T \cup U$ and $P(tu)$ is the marginal probability awarded to $tu@(T \cup U)$; if $t \not\sim u$ then $P(tu)$ is taken to be 0.

A *conditional probability function* induced by P is a function R from pairs of assignments of subsets of V to non-negative real numbers which satisfies (for each $t@T \subseteq V, u@U \subseteq V$):

$$R(t|u)P(u) = P(tu),$$

$$\sum_{t@T} R(t|u) = 1.$$

Note that $R(t|u)$ is not uniquely determined by P when $P(u) = 0$. If $P(u) \neq 0$ and the first condition holds, then the second condition, $\sum_{t@T} R(t|u) = 1$, also holds. Again, R is often thought of as an extension of P and is usually denoted by the same letter P.

Consider an example. Take a set of variables $V = \{A, B\}$, where A signifies *age of vehicle* taking possible values *less than 3 years*, *3-10 years* and *greater than 10 years*, and B signifies *breakdown in the last year* taking possible values *yes* and *no*. An assignment $b@B$ is of the form $B = yes$ or $B = no$. The assignments $a@A$ are most naturally written $A < 3, 3 \leq A \leq 10$ and $A > 10$.

According to the above definition a probability function P on V assigns a non-negative real number to each assignment of the form ab where $a@A$ and $b@B$, and these numbers must sum to 1. For instance,

$$P(A < 3 \cdot B = \mathit{yes}) = 0.05$$

$$P(A < 3 \cdot B = \mathit{no}) = 0.1$$

$$P(3 \leq A \leq 10 \cdot B = \mathit{yes}) = 0.2$$

$$P(3 \leq A \leq 10 \cdot B = \mathit{no}) = 0.2$$

$$P(A > 10 \cdot B = \mathit{yes}) = 0.35$$

$$P(A > 10 \cdot B = \mathit{no}) = 0.1.$$

This function P can be extended to assignments of subsets of V, yielding $P(A > 10) = P(A > 10 \cdot B = \mathit{yes}) + P(A > 10 \cdot B = \mathit{no}) = 0.35 + 0.1 = 0.45$ for example, and to conjunctions of assignments in which case inconsistent assignments are awarded probability 0, e.g., $P(B = \mathit{yes} \cdot B = \mathit{no}) = 0$. The function P can then be extended to yield conditional probabilities and, in this example, the probability of a breakdown conditional on age greater than 10 years, $P(B = \mathit{yes}|A > 10)$, is $P(A > 10 \cdot B = \mathit{yes})/P(A > 10) = 0.35/0.45 \approx 0.78$.

§3
Events

While the definition of probability over assignments to variables is straightforward, simplicity is gained at the expense of generality. By moving from variables to abstract events we can capture generality. The main definition proceeds as follows.[1]

Abstract *events* are construed as subsets of an *outcome space* Ω, which represents the possible outcomes of an experiment or observation. For example, if the age of a vehicle were observed, the outcome space might be $\Omega = \{0, 1, 2, \ldots\}$, and $\{0, 1, 2\} \subseteq \Omega$ represents the event that the vehicle's age is less than three years.

An *event space* $\mathcal{F}$ is a set of subsets of Ω. $\mathcal{F}$ is a *field* if it contains Ω and is closed under the formation of complements and finite unions; it is a *σ-field* if it is also closed under the formation of countable unions.

A *probability function* is a function P from a field $\mathcal{F}$ to the non-negative real numbers that satisfies *countable additivity*:

- if $E_1, E_2, \ldots \in \mathcal{F}$ partition Ω (i.e., $E_i \cap E_j = \emptyset$ for $i \neq j$ and $\bigcup_{i=1}^{\infty} E_i = \Omega$) then $\sum_{i=1}^{\infty} P(E_i) = 1$.

In particular, $P(\Omega) = 1$. The triple $(\Omega, \mathcal{F}, P)$ is called a *probability space*.

The variable framework is captured by letting Ω contain all assignments to V and taking $\mathcal{F}$ to be the set of all subsets of Ω, which corresponds to the set of disjunctions of assignments to V. Given variable $A \in V$, the function that maps $v@V$ to the value that v assigns to A is called a *simple random variable* in the event framework.

[1] Billingsley (1979) provides a good introduction to the theory behind this approach.

§4
Sentences

Logicians tend to define probability over logical languages (see, e.g., Paris, 1994). The simplest such framework is based around the propositional calculus, as follows.

A *propositional variable* is a variable which takes two possible values, *true* or *false*. A set $\mathcal{L}$ of propositional variables constitutes a *propositional language*. The sentences $S\mathcal{L}$ of $\mathcal{L}$ include the propositional variables, together with the *negation* $\neg\theta$ of each sentence $\theta \in S\mathcal{L}$ (which is true iff θ is false) and each *implication* of the form $\theta \to \varphi$ for $\theta, \varphi \in S\mathcal{L}$ (which is true iff θ is false or both θ and φ are true). The *conjunction* $\theta \wedge \varphi$ is defined to be $\neg(\theta \to \neg\varphi)$ and is true iff both θ and φ are true; the *disjunction* $\theta \vee \varphi$ is defined to be $\neg\theta \to \varphi$ and is true iff either θ or φ are true. An assignment l of values to $\mathcal{L}$ *models* sentence θ, written $l \models \theta$, if θ is true under l. A sentence θ is a *tautology*, written $\models \theta$, if it is true whatever the values of the propositional variables in θ, i.e., if each assignment to $\mathcal{L}$ models θ.

A *probability function* is then a function P from a set $S\mathcal{L}$ of sentences to the non-negative real numbers that satisfies *additivity*:

- if $\theta_1, \ldots, \theta_n \in S\mathcal{L}$ satisfy $\models \neg(\theta_i \wedge \theta_j)$ for $i \neq j$ and $\models \theta_1 \vee \cdots \vee \theta_n$ then $\sum_{i=1}^{n} P(\theta_i) = 1$.

If the language $\mathcal{L}$ is finite then the sentence framework can be mapped to the variable framework. $V = \mathcal{L}$ is a finite set of variables each of which takes finitely many values. A sentence $\theta \in SV$ can be identified with the set of assignments v of values to V which model θ. P thus maps sets of assignments and, in particular, individual assignments, to real numbers. P is additive because of additivity on sentences. Hence P induces a probability function over assignments to V.

The sentence framework can also be mapped to the event framework. Let Ω contain all assignments to L, and let $\mathcal{F}$ be the field of sets of the form $\{l : l \models \theta\}$ for $\theta \in S\mathcal{L}$.[2] By defining $P(\{l : l \models \theta\}) = P(\theta)$ we get a probability function.[3]

Part II
Interpretations of Probability

§5
Interpretations and Distinctions

The definitions of probability given in Part I are purely formal. In order to apply the formal concept of probability we need to know how probability is to be interpreted. The standard interpretations of probability will be presented in the next few sections.[4] These interpretations can be categorised according to the stances they take on three key distinctions:

[2] These sets are called *cylinder sets* when $\mathcal{L}$ is infinite—see Billingsley (1979, p. 27).

[3] This depends on the fact that every probability function on the field of cylinders which is *finitely additive* (i.e., which satisfies $\sum_{i=1}^{n} P(E_i) = 1$ for partition $E_1, \ldots, E_n$ of Ω) is also countably additive. See Billingsley (1979, Theorem 2.3).

[4] For a more detailed exposition of the interpretations see Gillies (2000).

Single-Case / Repeatable: A variable is *single-case* (or *token-level*) if it can only be assigned a value once. It is *repeatable* (or *repeatably instantiatable* or *type-level*) if it can be assigned values more than once. For example, variable A standing for *age of car with registration AB01 CDE on January 1st 2010* is single-case because it can only ever take one value (assuming the car in question exists). If, however, A stands for *age of vehicles selected at random in London in 2010* then A is repeatable: it gets reassigned a value each time a new vehicle is selected.[5]

Mental / Physical: Probabilities are *mental*—or *epistemological* (Gillies, 2000) or *personalist*—if they are interpreted as features of an agent's mental state, otherwise they are *physical*—or *aleatory* (Hacking, 1975).

Subjective / Objective: Probabilities are *subjective* (or *agent-relative*) if two agents with the same evidence can disagree as to a probability value and yet neither of them be wrong. Otherwise they are *objective*.[6]

There are four main interpretations of probability: the frequency theory (§6), the propensity theory (§7), chance (§8) and Bayesianism (§9).

§6
Frequency

The *Frequency* interpretation of probability was propounded by Venn (1866) and Reichenbach (1935) and developed in detail in von Mises (1928) and von Mises (1964). Von Mises' theory can be formulated in our framework as follows. Given a set V of repeatable variables one can repeatedly determine the values of the variables in V and write down the observations as assignments to V. For example, one could repeatedly select cars and determine their age and whether they broke down in the last year, writing down $A < 3 \cdot B = no, A < 3 \cdot B = yes, A > 10 \cdot B = yes$, and so on. Under the assumption that this process of measurement can be repeated *ad infinitum*, we generate an infinite sequence of assignments $\mathcal{V} = (v_1, v_2, v_3, \ldots)$ called a *collective*.

Let $|v|^n_{\mathcal{V}}$ be the number of times assignment v occurs in the first n places of $\mathcal{V}$, and let $Freq^n_{\mathcal{V}}(v)$ be the frequency of v in the first n places of $\mathcal{V}$, i.e.,

$$Freq^n_{\mathcal{V}}(v) = \frac{|v|^n_{\mathcal{V}}}{n}.$$

Von Mises noted two things. First, these frequencies tend to stabilise as the number n of observations increases. Von Mises hypothesised that

Axiom of Convergence: $Freq^n_{\mathcal{V}}(v)$ tends to a fixed limit as $n \longrightarrow \infty$, denoted by $Freq_{\mathcal{V}}(v)$.

[5]'Single-case variable' is clearly an oxymoron because the value of a single-case variable does not vary. The value of a single-case variable may not be known, however, and one can still think of the variable as taking a range of possible values.

[6]Warning: some authors, such as Popper (1983, §3.3) and Gillies (2000, p. 20), use the term 'objective' for what I call 'physical'. However, their terminology has the awkward consequence that the interpretation of probability commonly known as 'objective Bayesianism' (described in Part III) does not get classed as 'objective'.

Second, gambling systems tend to be ineffective. A gambling system can be thought of as function for selecting places in the sequence of observations on which to bet, on the basis of past observations. Thus a *place selection* is a function $f(v_1, \ldots, v_n) \in 0, 1$, such that if $f(v_1, \ldots, v_n) = 0$ then no bet is to be placed on the $n+1$-st observation and if $f(v_1, \ldots, v_n) = 1$ then a bet is to be placed on the $n+1$-st observation. So betting according to a place selection gives rise to a sub-collective $\mathcal{V}_f$ of $\mathcal{V}$ consisting of the places of $\mathcal{V}$ on which bets are placed. In practice we can only use a place selection function if it is simple enough for us to compute its values: if we cannot decide whether $f(v_1, \ldots, v_n)$ is 0 or 1 then it is of no use as a gambling system. According to Church's thesis a function is computable if it belongs to the class of functions known as *recursive* functions (Church, 1936). Accordingly we define a *gambling system* to be a recursive place selection. A gambling system is said to be effective if we are able to make money in the long run when we place bets according to the gambling system. Assuming that stakes are set according to frequencies of $\mathcal{V}$, a gambling system f can only be effective if the frequencies of $\mathcal{V}_f$ differ to those of $\mathcal{V}$: if $\mathit{Freq}_{\mathcal{V}_f}(v) > \mathit{Freq}_{\mathcal{V}}(v)$ then betting *on* v will be profitable in the long run; if $\mathit{Freq}_{\mathcal{V}_f}(v) < \mathit{Freq}_{\mathcal{V}}(v)$ then betting *against* v will be profitable. We can then explicate von Mises' second observation as follows:

Axiom of Randomness: Gambling systems are ineffective: if $\mathcal{V}_f$ is determined by a recursive place selection f, then for each v, $\mathit{Freq}_{\mathcal{V}_f}(v) = \mathit{Freq}_{\mathcal{V}}(v)$.

Given a collective $\mathcal{V}$ we can then define—following von Mises—the probability of v to be the frequency of v in $\mathcal{V}$:

$$P(v) \stackrel{\text{df}}{=} \mathit{Freq}_{\mathcal{V}}(v).$$

Clearly $\mathit{Freq}_{\mathcal{V}}(v) \geq 0$. Moreover $\sum_{v@V} |v|^n_{\mathcal{V}} = n$ so $\sum_{v@V} \mathit{Freq}^n_{\mathcal{V}}(v) = 1$ and, taking limits, $\sum_{v@V} \mathit{Freq}_{\mathcal{V}}(v) = 1$. Thus P is indeed a well-defined probability function.

Suppose we have a statement involving probability function P on V. If we also have a collective $\mathcal{V}$ on V then we can interpret the statement to be saying something about the frequencies of $\mathcal{V}$, and as being true or false according to whether the corresponding statement about frequencies is true or false respectively. This is the frequency interpretation of probability. The variables in question are repeatable, not single-case, and the interpretation is physical, relative to a collective of potential observations, not to the mental state of an agent. The interpretation is objective, not subjective, in the sense that once the collective is fixed then so too are the probabilities: if two agents disagree as to what the probabilities are, then at most one of the agents is right.

§7
Propensity

Karl Popper initially adopted a version of von Mises' frequency interpretation (Popper, 1934, Chapter VIII), but later, with the ultimate goal of formulating an interpretation of probability applicable to single-case variables, developed what is called the *propensity* interpretation of probability (Popper, 1959; Popper, 1983, Part II). The propensity theory can be thought of as the frequency theory together with the following law:[7]

Axiom of Independence: If collectives $\mathcal{V}_1$ and $\mathcal{V}_2$ on V are generated by the same repeatable experiment (or repeatable conditions) then for all assignments v to V, $Freq_{\mathcal{V}_1}(v) = Freq_{\mathcal{V}_2}(v)$.

In other words frequency, and hence probability, attaches to a repeatable experiment rather than a collective, in the sense that frequencies do not vary with collectives generated by the same repeatable experiment. The repeatable experiment is said to have a propensity for generating the corresponding frequency distribution.

In fact, despite Popper's intentions, the propensity theory interprets probability defined over repeatable variables, not single-case variables. If, for example, V consists of repeatable variables A and B, where A stands for *age of vehicles selected at random in London in 2010* and B stands for *breakdown in the last year of vehicles selected at random in London in 2010*, then V determines a repeatable experiment, namely the selection of vehicles at random in London in 2010, and thus there is a natural propensity interpretation. Suppose, on the other hand, that V contains single-case variables A and B, standing for *age of car with registration AB01 CDE on January 1st 2010* and *breakdown in last year of car with registration AB01 CDE on January 1st 2010.* Then V defines an experiment, namely the selection of car AB01 CDE on January 1st 2010, but this experiment is not repeatable and does not generate a collective—it is a single case. The car in question might be selected by several different repeatable experiments, but these repeatable experiments need not yield the same frequency for an assignment v, and thus the probability of v is not determined by V. (This is known as the *reference class problem*: we do not know from the specification of the single case how to uniquely determine a repeatable experiment which will fix probabilities.) In sum, the propensity theory is, like the frequency theory, an objective, physical interpretation of probability over repeatable variables.

§8
Chance

The question remains as to whether one can develop a viable objective interpretation of probability over single-case variables—such a concept of probability is often called *chance*.[8] We saw that frequencies are defined relative to a collective and propensities are defined relative to a repeatable experiment; however, a single-case variable does not determine a unique collective or repeatable experiment and so neither approach allows us to attach probabilities directly to single-case variables. What then does fix the chances of a single-case variable? The view finally adopted by Popper was that the 'whole physical situation' determines probabilities (Popper, 1990, p. 17). The physical situation might be

[7] Popper (1983, pp. 290 and 355). It is important to stress that the axioms of this section and the last had a different status for Popper than they did for von Mises. Von Mises used the frequency axioms as part of an operationalist definition of probability, but Popper was not an operationalist. See Gillies (2000, Chapter 7) on this point. Gillies also argues in favour of a propensity interpretation.

[8] Note that some authors use 'propensity' to cover a physical chance interpretation as well as the propensity interpretation discussed above.

thought of as 'the complete situation of the universe (or the light-cone) at the time' (Miller, 1994, p. 186), the complete history of the world up till the time in question (Lewis, 1980, p. 99),[9] or 'a complete set of (nomically and/or causally) relevant conditions ... which happens to be instantiated in that world at that time' (Fetzer, 1982, p. 195). Thus the chance, on January 1st 2010, of car with registration AB01 CDE breaking down in the subsequent year, is fixed by the state of the universe at that date, or its entire history up till that date, or all the relevant conditions instantiated at that date. However the chance-fixing 'complete situation' is delineated, these three approaches associate a unique chance-fixer with a given single-case variable. (In contrast, the frequency / propensity theories do not associate a unique collective / repeatable experiment with a given single-case variable.) Hence we can interpret the probability of an assignment to the single-case variable as the chance of the assignment holding, as determined by its chance-fixer.

Further explanation is required as to how one can measure probabilities under the chance interpretation. Popper's line is this: if the chance-fixer is a set of relevant conditions and these conditions are repeatable, then the conditions determine a propensity and that can be used to measure the chance (Popper, 1990, p. 17). Thus if the set of conditions relevant to car AB01 CDE breaking down that hold on January 1st 2010 also hold for other cars at other times, then the chance of AB01 CDE breaking down in the next year can be equated with the frequency with which cars satisfying the same set of conditions break down in the subsequent year. The difficulty with this view is that it is hard to determine all the chance-fixing relevant conditions, and there is no guarantee that enough individuals will satisfy this set of conditions for the corresponding frequency to be estimable.

§9
Bayesianism

The *Bayesian* interpretation of probability also deals with probability functions defined over single-case variables. But in this case the interpretation is mental rather than physical: probabilities are interpreted as an agent's rational degrees of belief.[10] Thus for an agent, $P(B = yes) = q$ if and only if the agent believes that $B = yes$ to degree q and this ascription of degree of belief is rational in the sense outlined below. An agent's degrees of belief are construed as a guide to her actions: she believes $B = yes$ to degree q if and only if she is prepared to place a bet of qS on $B = yes$, with return S if $B = yes$ turns out to be true. Here S is an unknown stake, which may be positive or negative, and q is called a *betting quotient*. An agent's *belief function* is the function that maps an assignment to the agent's degree of belief in that assignment.

An agent's betting quotients are called *coherent* if one cannot choose stakes for her bets that force her to lose money whatever happens. (Such a set of stakes is called a *Dutch book*.) It is not hard to see that a coherent belief function is a probability function. First $q \geq 0$, for otherwise one can set S to be negative and

[9] See §§10, 20.

[10] This interpretation was developed in Ramsey (1926) and de Finetti (1937). See Howson and Urbach (1989) and Earman (1992) for recent expositions.

the agent will lose whatever happens: she will lose $qS > 0$ if the assignment on which she is betting turns out to be false and will lose $(q-1)S > 0$ if it turns out to be true. Moreover $\sum_{v@V} q_v = 1$, where q_v is the betting quotient on assignment v, for otherwise if $\sum_v q_v > 1$ we can set each $S_v = S > 0$ and the agent will lose $(\sum_v q_v - 1)S > 0$ (since exactly one of the v will turn out true), and if $\sum_v q_v < 1$ we can set each $S_v = S < 0$ to ensure positive loss.

Coherence is taken to be a necessary condition for rationality. For an agent's degrees of belief to be rational they must be coherent, and hence they must be probabilities. *Subjective Bayesianism* is the view that coherence is also sufficient for rationality, so that an agent's belief function is rational if and only if it is a probability function. This interpretation of probability is subjective because it depends on the agent as to whether $P(v) = q$. Different agents can choose different probabilities for v and their belief functions will be equally rational. *Objective Bayesianism*, discussed in detail in Part III, imposes further rationality constraints on degrees of belief—not just coherence. Very often objective Bayesianism constrains degree of belief in such a way that only one value for $P(v)$ is deemed rational on the basis of an agent's evidence. Thus, objective Bayesian probability varies as evidence varies but two agents with the same evidence often adopt the same probabilities as their rational degrees of belief.[11]

Many subjective Bayesians claim that an agent should update her degrees of belief by *Bayesian conditionalisation*: her new degrees of belief should be her old degrees of belief conditional on new evidence, $P_{t+1}(v) = P_t(v|u)$ where u represents the evidence that the agent has learned between time t and time $t+1$. In cases where $P_t(v|u)$ is harder to quantify than $P_t(u|v)$ and $P_t(v)$ this conditional probability may be calculated using *Bayes' theorem*: $P(v|u) = P(u|v)P(v)/P(u)$, which holds for any probability function P. Note that Bayesian conditionalisation is more appropriate as a constraint on subjective Bayesian updating than on objective Bayesian updating, because it disagrees with the usual principles of objective Bayesianism (Williamson, 2008b). 'Bayesianism' is variously used to refer to the Bayesian interpretation of probability, the endorsement of Bayesian conditionalisation or the use of Bayes' theorem.

§10
Chance as Ultimate Belief

The question still remains as to whether one can develop a viable notion of chance, i.e., an objective single-case interpretation of probability. While the Bayesian interpretations are single-case, they either define probability relative to the whimsy of an agent (subjective Bayesianism) or relative to an agent's evidence (objective Bayesianism). Is there a probability of my car breaking down in the next year, where this probability does not depend on me or my evidence?

Bayesians typically have two ways of tackling this question.

Subjective Bayesians tend to argue that although degrees of belief may initially vary widely from agent to agent, if agents update their degrees of belief by Bayesian conditionalisation then their degrees of belief will converge in the long

[11] Objective Bayesian degrees of belief are uniquely determined on a finite set of variables; on infinite domains subjectivity can creep in (§19).

run: chances are these long run degrees of belief. Bruno de Finetti developed such an argument to explain the apparent existence of physical probabilities (de Finetti, 1937; Gillies, 2000, pp. 69–83). He showed that prior degrees of beliefs converge to frequencies under the assumption of *exchangeability*: given an infinite sequence of single-case variables $A_1, A_2, \ldots$ which take the same possible values, an agent's degrees of belief are *exchangeable* if the degree of belief $P(v)$ she gives to assignment v to a finite subset of variables depends only on the values in v and not the variables in v—for example $P(a_1^1 a_2^0 a_3^1) = P(a_3^0 a_4^1 a_5^1)$ since both assignments assign two 1s and one 0. Suppose the actual observed assignments are $a_1, a_2, \ldots$ and let $\mathcal{V}$ be the collective of such values (which can be thought of as arising from a single repeatable variable A). De Finetti showed that $P(a_n|a_1 \cdots a_{n-1}) \longrightarrow \mathit{Freq}_{\mathcal{V}}(a)$ as $n \longrightarrow \infty$, where a is the assignment to A of the value that occurs in a_n. The chance of a_n is then identified with $\mathit{Freq}_{\mathcal{V}}(a)$. The trouble with de Finetti's account is that since degrees of belief are subjective there is no reason to suppose exchangeability holds. Moreover, a single-case variable A_n can occur in several sequences of variables, each with a different frequency distribution (the reference class problem again), in which case the chance distribution of A_n is ill-defined. Haim Gaifman and Marc Snir took a slightly different approach, showing that as long as agents give probability 0 to the same assignments and the evidence that they observe is unrestricted, then their degrees of belief must converge (Gaifman and Snir, 1982, §2). Again, the problem here is that there is no reason to suppose that agents will give probability 0 to the same assignments. One might try to provide such a guarantee by bolstering subjective Bayesianism with a rationality constraint that says that agents must be *undogmatic*, i.e., they must only give probability 0 to logically impossible assignments. But this is not a feasible strategy in general, since this constraint is inconsistent with the constraint that degrees of belief be probabilities: in the more general event or sentence frameworks the laws of probability force some logical possibilities to be given probability 0.[12]

Objective Bayesians have another recourse open to them: objective Bayesian probability is fixed by an agent's evidence, and one can argue that chances are those degrees of belief fixed by some suitable all-encompassing evidence. Thus the problem of producing a well-defined notion of chance is reducible to that of developing an objective Bayesian interpretation of probability. I shall call this the *ultimate belief* notion of chance to distinguish it from physical notions such as Popper's (§8), and discuss this approach in §20.

§11
Applying Probability

In sum, there are four key interpretations of probability: frequency and propensity interpret probability over repeatable variables while chance and the Bayesian interpretation deal with single-case variables; frequency and propensity are physical interpretations while Bayesianism is mental and chance can be either mental or physical; all the interpretations are objective apart from Bayesianism which can be subjective or objective.

Having chosen an interpretation of probability, one can use the probability

[12]See Gaifman and Snir (1982, Theorem 3.7), for example.

calculus to draw conclusions about the world. Typically, having made an observation $u@U \subseteq V$, one determines the conditional probability $P(t|u)$ to tell us something about $t@T \subseteq (V \backslash U)$: a frequency, propensity, chance or appropriate degree of belief.

Part III
Objective Bayesianism

§12
Subjective and Objective Bayesianism

In Part II we saw that probabilities can either be interpreted physically—as frequencies, propensities or physical chances—or they can be interpreted mentally, with Bayesians arguing that an agent's degrees of belief ought to satisfy the axioms of probability. Some Bayesians are strict subjectivists, holding that there are no rational constraints on degrees of belief other than the requirement that they be probabilities (de Finetti, 1937). Thus subjective Bayesians maintain that one may give probability 0—or indeed any value between 0 and 1—to a coin toss yielding heads, even if one knows that the coin is symmetrical and has yielded heads in roughly half of all its previous tosses. The chief criticism of strict subjectivism is that practical applications of probability tend to demand more objectivity; in science some beliefs are considered more rational than others on the basis of available evidence. This motivates an alternative position, objective Bayesianism, which posits further constraints on degrees of belief, and which would only deem the agent to be rational in this case if she gave a probability of a half to the toss yielding heads (Jaynes, 1988).

Objective Bayesianism holds that the probability of u is the degree to which an agent ought to believe u and that this degree is more or less objectively determined by the agent's evidence. Versions of this view were put forward by Bernoulli (1713); Laplace (1814) and Keynes (1921). More recently Jaynes claimed that an agent's probabilities ought to satisfy constraints imposed by evidence but otherwise ought to be as non-committal as possible. Moreover, Jaynes argued, this principle could be explicated using Shannon's information theory (Shannon, 1948): the agent's probability function should be that probability function, from all those that satisfy constraints imposed by evidence, that maximises entropy (Jaynes, 1957). This has become known as *the Maximum Entropy Principle* and has been taken to be the foundation of the objective Bayesian interpretation of probability by its proponents (Rosenkrantz, 1977; Jaynes, 2003).

In the next section, I shall sketch my own version of objective Bayesianism. This version is discussed in detail in chapter 4 of Williamson (2005a). In subsequent sections we shall examine a range of important challenges that face the objective Bayesian interpretation of probability.

§13
Objective Bayesianism Outlined

While Bayesianism requires that degrees of belief respect the axioms of probability, objective Bayesianism imposes two further norms. An empirical norm requires that an agent's degrees of belief be calibrated with her evidence, while a logical norm holds that where degrees of belief are underdetermined by evidence, they should be as equivocal as possible:

Empirical: An agent's empirical evidence should constrain her degrees of belief. Thus if one knows that a coin is symmetrical and has yielded heads roughly half the time, then one's degree of belief that it will yield heads on the next throw should be roughly $\frac{1}{2}$.

Logical: An agent's degrees of belief should also be fixed by her lack of evidence. If the agent knows nothing about an experiment except that it has two possible outcomes, then she should award degree of belief $\frac{1}{2}$ to each outcome.

Jakob Bernoulli pointed out that where they conflict, the empirical norm should override the logical norm:

> three ships set sail from port; after some time it is announced that one of them suffered shipwreck; which one is guessed to be the one that was destroyed? If I considered merely the number of ships, I would conclude that the misfortune could have happened to each of them with equal chance; but because I remember that one of them had been eaten away by rot and old age more than the others, had been badly equipped with masts and sails, and had been commanded by a new and inexperienced captain, I consider that this ship, more probably than the others, was the one to perish (Bernoulli, 1713, §IV.II)

One can prioritise the empirical norm over the logical norm by insisting that:

Empirical: An agent's degrees of belief, represented by probability function $P_{\mathcal{E}}$, should satisfy any constraints imposed by her evidence $\mathcal{E}$.

Logical: The agent's belief function $P_{\mathcal{E}}$ should otherwise be as non-committal as possible.

The empirical norm can be explicated as follows. Evidence $\mathcal{E}$ might contain a number of considerations that bear on a degree of belief: the symmetry of a penny might incline one to degree of belief $\frac{1}{2}$ in heads, past performance (say 47 heads in a hundred past tosses) may incline one to degree of belief 0.47, the mint may report an estimate of the frequency of heads on its pennies to be 0.45, and so on. These considerations may be thought of as conflicting reports as to the probability of heads. Intuitively, any individual report, say 0.47, is compatible with the evidence, and indeed intermediary degrees of belief such as 0.48 seem reasonable. On the other hand, a degree of belief that falls outside the range of reports, say 0.9, does not seem warranted by the evidence. Thus evidence constrains degree of belief to lie in the smallest closed interval that contains all the reports.

As mentioned in §12, the logical norm is explicated using the Maximum Entropy Principle: entropy is a measure of the lack of commitment of a probability function, so $P_{\mathcal{E}}$ should be the probability function, out of all those that satisfy constraints imposed by $\mathcal{E}$, that has maximum entropy. Justifications of the Maximum Entropy Principle are well known—see Jaynes (2003), Paris (1994) or Paris and Vencovská (2001) for example.

We can thus put the two norms on a more formal footing. Given a domain V of finitely many variables, each of which takes finitely many values, an agent with evidence $\mathcal{E}$ should adopt as her belief function the probability function $P_{\mathcal{E}}$ on V determined as follows:

Empirical: $P_{\mathcal{E}}$ should satisfy any constraints imposed by her evidence $\mathcal{E}$: $P_{\mathcal{E}}$ should lie in the smallest closed convex set $\mathbb{E}$ of probability functions containing those probability functions that are compatible with the reports in $\mathcal{E}$.[13]

Logical: $P_{\mathcal{E}}$ should otherwise be as non-committal as possible: $P_{\mathcal{E}}$ should be a member of $\mathbb{E}$ that maximises entropy $H(P) = -\sum_{v@V} P(v) \log P(v)$.

It turns out that there is a unique entropy maximiser on a closed convex set of probability functions: the degrees of belief $P_{\mathcal{E}}$ that an agent should adopt are uniquely determined by her evidence $\mathcal{E}$. Thus on a finite domain there is no room for subjective choice of degrees of belief.

§14
Challenges

While objective Bayesianism is popular amongst practitioners—e.g., in statistics, artificial intelligence, physics and engineering—it has not been widely accepted by philosophers, however, largely because there are a number of perceived problems with the interpretation. Several of these problems have in fact already been resolved, but other challenges remain. In the remainder of this part of the paper we shall explore the key challenges and assess the prospects of objective Bayesianism.

In §15 we shall see that one challenge is to motivate the adoption of a logical norm. Objective Bayesianism has also been criticised for being language dependent (§16) and for being impractical from a computational point of view (§17). Handling qualitative evidence poses a significant challenge (§18), as does extending objective Bayesianism to infinite event or sentence frameworks (§19). The question of whether objective Bayesianism can be used to provide an interpretation of objective chance is explored in §20, while §21 considers the application of objective Bayesianism to providing semantics for probability logic.

Jaynes points out that the Maximum Entropy Principle is a powerful tool but warns:

> Of course, it is as true in probability theory as in carpentry that introduction of more powerful tools brings with it the obligation

[13]See Williamson (2005a, §5.3) for more detailed discussion of this norm. There it is argued that $\mathbb{E}$ is constrained not only by quantitative evidence of physical probability but also evidence of qualitative relations between variables such as causal relations. See §18 on this point.

> to exercise a higher level of understanding and judgement in using them. If you give a carpenter a fancy new power tool, he *may* use it to turn out more precise work in greater quantity; or he may just cut off his thumb with it. It depends on the carpenter (Jaynes, 1979, pp. 40–41 of the original 1978 lecture).

§15
Motivation

The first key question concerns the motivation behind objective Bayesianism. Recall that in §12 objective Bayesianism was motivated by the need for objective probabilities in science. Many Bayesians accept this desideratum and indeed accept the empirical norm (so that degrees of belief are constrained by evidence of frequencies, symmetries, etc.) but do not go as far as admitting a logical norm. The ensuing position, according to which degrees of belief reflect evidence but need not be maximally non-committal, is sometimes called *empirically-based subjective probability*. It yields degrees of belief that are more objective (i.e., more highly constrained) than those of strictly subjective Bayesianism, yet not as objective as those of objective Bayesianism—there is generally still some room for subjective choice of degrees of belief. The key question is thus: what grounds are there for going beyond empirically-based subjective probability and adopting objective Bayesianism?

Current justifications of the logical norm fail to address this question. Jaynes' original justification of the Maximum Entropy Principle ran like this: *given that degrees of belief ought to be maximally non-committal*, Shannon's information theory shows us that they are entropy-maximising probabilities (Jaynes, 1957). This type of justification assumes from the outset that some kind of logical norm is desired. On the other hand, axiomatic derivations of the Maximum Entropy Principle take the following form: *given that we need a procedure for objectively determining degrees of belief from evidence*, and given various desiderata that such a procedure should satisfy, that procedure must be entropy maximisation (Paris and Vencovská, 1990; Paris, 1994; Paris and Vencovská, 2001). This type of justification takes objectivity of rational degrees of belief for granted. Thus the challenge is to augment current justifications, perhaps by motivating non-committal degrees of belief or by motivating the strong objectivity of objective Bayesianism as opposed to the partial objectivity yielded by empirically-based subjective probability.

One possible approach is to argue that empirically-based subjective probability is *not objective enough* for many applications of probability. Many applications of probability follow a Bayesian statistical methodology: produce a *prior* probability function P_t, collect some evidence u, and draw predictions using the *posterior* probability function $P_{t+1}(v) = P_t(v|u)$. Now the prior function is determined before empirical evidence is available; this is matter of subjective choice for empirically-based subjectivists. However, the ensuing conclusions and predictions may be sensitive to this initial choice, rendering them subjective too. Yet such relativism is anathema in science: a disagreement between agents about a hypothesis should be arbitrated by evidence; it should be a fact of the matter, not mere whim, as to whether the evidence confirms the hypothesis.

That argument is rather inconclusive however. The proponent of empirically-based subjective probability can counter that scientists have simply over-estimated the extent of objectivity in science, and that subjectivity needs to be made explicit. Even if one grants a need for objectivity, one could argue that it is a pragmatic need: it just makes science simpler. The objective Bayesian must accept that it cannot be empirical warrant that motivates the selection of a particular belief function from all those compatible with evidence, since all such belief functions are equally warranted by available empirical evidence. In the absence of any non-empirical justification for choosing a particular belief function, such a function can only be considered objective in a *conventional* sense. One can drive on the right or the left side of the road; but we must all do the same thing; by convention in the UK we choose the left. That does not mean that the left is objectively correct or most warranted—either side will do.

A second line of argument offers explicitly pragmatic reasons for selecting a particular belief function. If probabilities are subjective then measuring probabilities must involve elicitation of degrees of belief from agents. As developers of expert systems in AI have found, elicitation and the associated consistency-checking are prohibitively time-consuming tasks (the inability of elicitation to keep pace with the demand for expert systems is known as *Feigenbaum's bottleneck*). If a subjective approach is to be routinely applied throughout science it is clear that a similar bottleneck will be reached. On the other hand, if degrees of belief are objectively determined by evidence then elicitation is not required—degrees of belief are calculated by maximising entropy. Objective Bayesianism is thus to be preferred for reasons of efficiency.

Indeed many Bayesian statisticians now (often tacitly) appeal to non-committal objective priors rather than embark on a laborious process of introspection, elicitation or analysis of sensitivity of posterior to choice of prior.

A third motivating argument appeals to caution. In many applications of probability the risks attached to bold predictions that turn out wrong are high. For instance, a patient's symptoms may narrow her condition down to meningitis or 'flu, but there may be no empirical evidence—such as information about relative prevalence—to decide between the two. In this case, the risks associated with meningitis are so much higher than those associated with 'flu, that a non-committal belief function seems more appropriate as a basis for action than a belief function that gives the probability of meningitis to be zero, even though both are compatible with available information. (With a non-committal belief function one will not dismiss the possibility of meningitis, but if one gives meningitis probability zero one will disregard it.) High-risk applications thus favour cautious conclusions, non-committal degrees of belief and an objective Bayesian approach.

I argue in Williamson (2007b) that the appeal to caution is the most decisive motivation for objective Bayesianism, although pragmatic considerations play a part too.

§16
Language Dependence

The Maximum Entropy Principle has been criticised for being language or representation dependent: it has been argued that the principle awards the same event different probabilities depending on the way in which the problem domain is formulated.

John Maynard Keynes surveyed several purported examples of language dependence in his discussion of Laplace's Principle of Indifference (Keynes, 1921). This latter principle advocates assigning the same probability to each of a number of possible outcomes in the absence of any evidence which favours one outcome over the others. Keynes added the condition that the possible outcomes must be indivisible (Keynes, 1921, §4.21). The Maximum Entropy Principle makes the same recommendation in the absence of evidence and so inherits any language dependence of the Principle of Indifference.

A typical example of language dependence proceeds as follows (Halpern and Koller, 1995, §1). Suppose an agent's language can be represented by the propositional language $\mathcal{L} = \{C\}$ with just one propositional variable C which asserts that a particular book is colourful. The agent has no evidence and so by the Principle of Indifference (or equally by the Maximum Entropy Principle) assigns $P(C) = P(\neg C) = 1/2$. But now consider a second language $\mathcal{L}' = \{R, B, G\}$ where R signifies that the book is red, B that it is blue and G that it is green. An agent with no evidence will give $P(\pm R \wedge \pm B \wedge \pm G) = 1/8$. Now $\neg C$ is equivalent to $\neg R \wedge \neg B \wedge \neg G$, yet the former is given probability $\frac{1}{2}$ while the latter is given probability $\frac{1}{8}$. Thus the probability assignments of the Principle of Indifference and the Maximum Entropy Principle depend on choice of language.

Paris and Vencovská (1997) offer the following resolution. They argue that the Maximum Entropy Principle has been misapplied in this type of example: if an agent refines the propositional variable C into $R \vee B \vee G$ one should consider not $\mathcal{L}'$ but $\mathcal{L}'' = \{C, R, B, G\}$ and make the agent's evidence, namely $C \leftrightarrow R \vee B \vee G$, explicit. If we do that then the probability function on $\mathcal{L}''$ with maximum entropy, out of all those that satisfy the evidence (i.e., which assign $P(C \leftrightarrow R \vee B \vee G) = 1$), will yield a value $P(\neg C) = 1/2$. This is just the same value as that given by the Maximum Entropy Principle on $\mathcal{L}$ with no evidence. Thus there is no inconsistency.

This resolution is all well and good if we are concerned with a single agent who refines her language. But the original problem may be construed rather differently. If *two* agents have languages $\mathcal{L}$ and $\mathcal{L}'$ respectively, and no evidence, then they assign two different probabilities to what we know (but they don't know) is the same proposition. There is no getting round it: probabilities generated by the Maximum Entropy Principle depend on language as well as evidence.

Interestingly, language dependence in this latter multilateral sense is not confined to the Maximum Entropy Principle. As Halpern and Koller (1995) and Paris and Vencovská (1997) point out, there is no non-trivial principle for selecting rational degrees of belief which is language-independent in the multilateral sense. More precisely, suppose we want a principle that selects a set $\mathbb{O}_{\mathcal{E}}$ of probability functions that are optimally rational on the basis of an agent's evidence $\mathcal{E}$. If $\mathbb{O}_{\mathcal{E}} \subseteq \mathbb{E}$, i.e., if every optimally rational probability function must satisfy constraints imposed by $\mathcal{E}$, and if $\mathbb{O}_{\mathcal{E}}$ ignores irrelevant information inasmuch as $\mathbb{O}_{\mathcal{E}\cup\mathcal{E}'}(\theta) = \mathbb{O}_{\mathcal{E}}(\theta)$ whenever $\mathcal{E}'$ involves no propositional variables in sentence θ, then the only candidate for $\mathbb{O}_{\mathcal{E}}$ that is multilaterally language independent is $\mathbb{O}_{\mathcal{E}} = \mathbb{E}$ (Halpern and Koller, 1995, Theorem 3.10). Only empirically-based subjective probability is multilaterally language independent.

So much the better for empirically-based subjective probability and so much the worse for objective Bayesianism, one might think. But such an inference is too quick. It takes the desirability of multilateral language independence for granted. I argue in Williamson (2005a, Chapter 12) that an agent's language constitutes empirical evidence:[14] evidence of natural kinds, evidence concerning which variables are relevant to which, and perhaps even evidence of which partitions are amenable to the Principle of Indifference. For example, having dozens of words for snow in one's language says something about the environment in which one lives. Granted that language itself is a kind of evidence, and granted that an agent's degrees of belief should depend on her evidence, language independence becomes a rather dubious desideratum.

Note that while Howson (2001, p. 139) criticises the Principle of Indifference on account of its language dependence, the example he cites can be used to support the case *against* language independence as a desideratum. Howson considers two first-order languages with equality: $\mathcal{L}_1$ has just a unary predicate U while $\mathcal{L}_2$ has unary U together with two constants t_1 and t_2. The explicit evidence $\mathcal{E}$ is just 'there are exactly 2 individuals', while sentence θ is 'something has the property U'. $\mathcal{L}_1$ has three models of $\mathcal{E}$, which contain 0, 1 and 2 instances of U respectively, so $P(\theta) = 2/3$. In $\mathcal{L}_2$ individuals can be distinguished by constants and thus there are eight models of $\mathcal{E}$ (if constants can name the same individual), six of which satisfy θ so $P(\theta) = 3/4 \neq 2/3$. While this is a good example of language dependence, the question remains whether language dependence is a problem here. As Howson himself hints, $\mathcal{L}_1$ might be an appropriate language for talking about bosons, which are indistinguishable, while $\mathcal{L}_2$ is more suited to talk about classical particles, which are distinguishable and thus able to be named by constants. Hence choice of language $\mathcal{L}_2$ over $\mathcal{L}_1$ indicates distinguishability, while conversely choice of $\mathcal{L}_1$ over $\mathcal{L}_2$ indicates indistinguishability. In this example, then, language betokens implicit evidence. Of course all but the the most ardent subjectivists agree that an agent's degrees of belief ought to be influenced by her evidence. Therefore language independence becomes an inappropriate desideratum.

In sum, while the Principle of Indifference and the Maximum Entropy Principle have both been dismissed on the grounds of language dependence, it seems clear that some dependence on language is to be expected if degrees of belief are to adequately reflect implicit as well as explicit evidence. So much the better for objective Bayesianism, and so much the worse for empirically-based subjective probability which is language-invariant.

[14] Halpern and Koller (1995, §4) also suggest this tack, although they do not give their reasons. Interestingly, though, they do show in §5 that relaxing the notion of language independence leads naturally to an entropy-based approach.

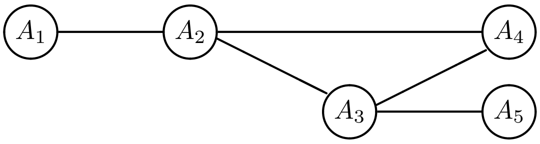

Figure 1: A constraint graph.

§17
Computation

There are important concerns regarding the application of objective Bayesianism. One would like to apply objective Bayesianism in artificial intelligence: when designing an artificial agent it would be very useful to have normative rules which prescribe how the agent's beliefs should change as it gathers information about its world. However, there has seemed to be little prospect of fulfilling this hope, for the following reason. Maximising entropy involves finding the parameters $P(v)$ that maximise the entropy expression, but the number of such parameters is exponential in the number of variables in the domain, thus the size of the entropy maximisation problem quickly gets out of hand as the size of the domain increases. Indeed Pearl (1988, p. 468) has influentially criticised maximum entropy methods on account of their computational difficulties.

The computational problem poses a serious challenge for objective Bayesianism. However, recent techniques for more efficient entropy maximisation have largely addressed this issue. While no technique offers efficient entropy maximisation in all circumstances (entropy maximisation is an NP-complete problem), techniques exist that offer efficiency in a wide range of natural circumstances. I shall sketch the theory of *objective Bayesian nets* here—this is developed in detail in Williamson (2005a, §§5.5–5.7) and Williamson (2005b).[15]

Given a set V of variables and some evidence $\mathcal{E}$ involving V which consists of a set of constraints on the agent's belief function P, one wants to find the probability function P, out of all those that satisfy the constraints in $\mathcal{E}$, that maximises entropy. This can be achieved via the following procedure. First form an undirected graph on vertices V by linking pairs of variables that occur in the same constraint with an edge. For example, if $V = \{A_1, A_2, A_3, A_4, A_5\}$ and $\mathcal{E}$ contains a constraint involving A_1 and A_2 (e.g., $P(a_2^1|a_1^0) = 0.9$), a constraint involving A_2, A_3 and A_4, a constraint involving A_3 and A_5 and a constraint involving just A_4, then the corresponding undirected constraint graph appears in Fig. 1. The undirected constraint graph has the following crucial property: if a set Z of variables separates $X \subseteq V$ from $Y \subseteq V$ in the graph then the maximum entropy function P will render X and Y probabilistically independent conditional on Z.

Next transform the undirected constraint graph into a directed constraint graph, Fig. 2 in the case of our example.[16] The independence property ensures that the directed constraint graph can be used as a graph in a *Bayesian net*

[15] Maximum entropy methods have recently been applied to natural language processing, and other techniques for entropy maximisation have been tailored to that context—see Della Pietra et al. (1997) for example.

[16] The algorithm for this transformation is given in Williamson (2005a, §5.7).

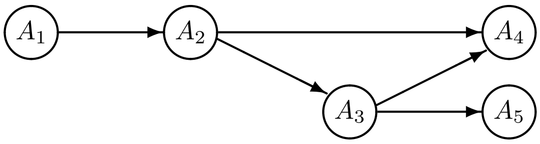

Figure 2: A directed constraint graph.

representation of the maximum entropy function P. A Bayesian net offers the opportunity of a more efficient representation of a probability function P: in order to determine P, one only needs to determine the parameters $P(a_i|par_i)$, i.e., the probability distribution of each variable conditional on its parents, rather than the parameters $P(v)$, i.e., the joint probability distribution over all the variables. Depending on the structure of the directed graph, there may be far fewer parameters in the Bayesian net representation. In the case of our example, if we suppose that each variable has two possible values then the Bayesian net representation requires 11 parameters rather than the 32 parameters $P(v)$ for each assignment v of values to V. For problems involving more variables the potential savings are very significant.

Roughly speaking, efficiency savings are greatest when each variable has few parents in the directed constraint graph, and this occurs when each constraint in $\mathcal{E}$ involves relatively few variables. Note that when dealing with large sets of variables it tends to be the case that while one might make a large number of observations, each observation involves relatively few variables. For example, one might use hospital data as empirical observations pertaining to a large number of health-related variables, each department of the hospital contributing some statistics; while there might be a large number of such statistics, each statistic is likely to involve relatively few variables, namely those variables that are relevant to the department in question; such observations would yield a sparse constraint graph and an efficient Bayesian net representation. Hence this method for reducing the complexity of entropy maximisation offers efficiency savings that are achievable in a wide range of natural situations.

A Bayesian net that represents the probability function produced by the Maximum Entropy Principle is called an *objective Bayesian net*. See Nagl et al. (2008) for an application for the objective Bayesian net approach to cancer prognosis and systems biology.

§18
Qualitative Knowledge

The Maximum Entropy Principle has been criticised for yielding the wrong results when the agent's evidence contains qualitative causal information (Pearl, 1988, p. 468; Hunter, 1989). Daniel Hunter gives the following example:

> The puzzle is this: Suppose that you are told that three individuals, Albert, Bill and Clyde, have been invited to a party. You know nothing about the propensity of any of these individuals to go to the party nor about any possible correlations among their actions. Using the obvious abbreviations, consider the eight-point space consisting

of the events $ABC, AB\bar{C}, A\bar{B}C$, etc. (conjunction of events is indicated by concatenation). With no constraints whatsoever on this space, MAXENT yields equal probabilities for the elements of this space. Thus $Prob(A) = Prob(B) = 0.5$ and $Prob(AB) = 0.25$, so A and B are independent. It is reasonable that A and B turn out to be independent, since there is no information that would cause one to revise one's probability for A upon learning what B does. However, suppose that the following information is presented: Clyde will call the host before the party to find out whether Al or Bill or both have accepted the invitation, and his decision to go to the party will be based on what he learns. Al and Bill, however, will have no information about whether or not Clyde will go to the party. Suppose, further, that we are told the probability that Clyde will go conditional on each combination of Al and Bill's going or not going. For the sake of specificity, suppose that these conditional probabilities are ...
$[P(C|AB) = 0.1, P(C|A\bar{B}) = 0.5, P(C|\bar{A}B) = 0.5, P(C|\bar{A}\bar{B}) = 0.8]$.

> When MAXENT is given these constraints ... A and B are no longer independent! But this seems wrong: the information about Clyde should not make A's and B's actions dependent (Hunter, 1989, p. 91)

But this counter-intuitive conclusion is attributable to a misapplication of the Maximum Entropy Principle. The conditional probabilities are allowed to constrain the entropy maximisation process but the knowledge that Al's and Bill's decisions are causes of Clyde's decision is simply ignored. This failure to consider the qualitative causal evidence leads to the counter-intuitive conclusion.

Keynes himself had stressed the importance of taking qualitative knowledge into account and the difficulties that ensue if qualitative information is ignored:

> Bernoulli's second axiom, that in reckoning a probability we must take everything into account, is easily forgotten in these cases of statistical probabilities. The statistical result is so attractive in its definiteness that it leads us to forget the more vague though more important considerations which may be, in a given particular case, within our knowledge (Keynes, 1921, p. 322).

Indeed, in the party example, the temptation is to consider only the definite probabilities and to ignore the important causal evidence.

The party example and Keynes' advice highlight an important challenge for objective Bayesianism. In order that objective Bayesianism can be applied, all evidence—*qualitative* as well as quantitative—must be taken into account. However, objective Bayesianism as outlined in §13 depends on evidence taking *quantitative* form: evidence must be explicated as a set of quantitative constraints on degrees of belief in order to narrow down a set of probability functions that satisfy those constraints. Thus the general challenge for objective Bayesianism is to show how qualitative evidence can be converted into precise quantitative constraints on degrees of belief.

To some extent this challenge has already been met. In the case where qualitative evidence takes the form of causal constraints, as in Hunter's party example above, I advocate a solution which exploits the following asymmetry of causality. Learning of the existence of a common cause of two events may warrant a change in the degrees of belief awarded to them: one may reason that if one event occurs, then this may well be because the common cause has occurred, in which case the other event is more likely—the two events become more dependent than previously thought. On the other hand, learning of the existence of a common effect would not warrant a change in degrees of belief: while the occurrence of one event may make the common effect more likely, this has no bearing on the other cause. This asymmetry motivates what I call the *Causal Irrelevance Principle*: if the agent's language contains a variable A that is known not to be a cause of any of the other variables, then her degrees of belief concerning these other variables should be the same as the degrees of belief she should adopt were she not to have A in her language (as long as any quantitative evidence involving A is compatible with those degrees of belief). The Causal Irrelevance Principle allows one to transfer qualitative causal evidence into quantitative constraints on degrees of belief—if domain $V = U \cup \{A\}$ then we have constraints of the form $P^V_{\restriction U} = P^U$, i.e., the agent's belief function defined on V, when restricted to U, should be the same as the belief function defined just on U. By applying the Causal Irrelevance Principle, qualitative causal evidence as well as quantitative information can be used to constrain the entropy maximisation process. It is not hard to see that use of the principle avoids counter-intuitive conclusions like those in Hunter's example: knowledge that Clyde's decision is a common effect of Al's and Bill's decision ensures that Al's and Bill's actions are probabilistically independent, as seems intuitively plausible. See Williamson (2005a, §5.8) for a more detailed analysis of this proposal.

Thus the challenge of handling qualitative evidence has been met in the case of causal evidence. Moreover, by treating logical influence analogously to causal influence one can handle qualitative logical evidence using the same strategy (Williamson, 2005a, Chapter 11). But the challenge has not yet been met in other cases of qualitative evidence. In particular, I claimed in §16 that choice of language implies evidence concerning the domain. Clearly work remains to be done to render such evidence explicit and quantitative, so that it can play a role in the entropy maximisation process.

There is another scenario in which the challenge is only beginning to be met. Some critics of the Maximum Entropy Principle argue that objective Bayesianism renders learning from experience impossible, as follows. The Maximum Entropy Principle will, in the absence of evidence linking them, render outcomes probabilistically independent. Thus observing outcomes will not change degrees of belief in unobserved outcomes if there is no evidence linking them: observing a million ravens, all black, will not shift the probability of the next raven being black from $\frac{1}{2}$ (which is the most non-committal value given only that there are two outcomes, black or not black). So, the argument concludes, there is no learning from experience. The problem with this argument is that we do have evidence that connects the outcomes—the qualitative evidence that we are repeatedly sampling ravens to check whether they are black—but this evidence is mistakenly being ignored in the application of the Maximum Entropy Principle. Qualitative evidence should be taken into account so that learning from experience becomes possible—but how? Carnap (1952) and Carnap (1971) addressed the problem, as have Paris and Vencovská (2003); Williamson (2007a) and Williamson (2008c) more recently. Broadly speaking, the idea behind this line of work is to take the maximally non-committal probability function to be one which permits learning from experience, as opposed to the maximum entropy probability function which does not. The difficulty with this approach is that it does genuinely seem to be the maximum entropy function that is most non-committal. An altogether different approach, developed in Williamson (2008b, §5), is to argue that learning from experience should result from the empirical norm rather than the logical norm: observing a million ravens, all black, does not merely impose the constraint that the agent should fully believe that *those* ravens are black—it also imposes the constraint that the agent should strongly if not fully believe that *other* (unobserved) ravens are also black. Then the agent's belief function should as usual be a function, from all those that satisfy these constraints, that has maximum entropy. This alternative approach places the problem of learning from experience firmly in the province of statistics rather than inductive logic.

§19
Infinite Domains

The Maximum Entropy Principle is most naturally defined on a finite domain—for example, a space of finitely many variables each of which takes finitely many values, as in §2. The question thus arises as to whether one can extend the applicability of objective Bayesianism to infinite domains. In the variable framework, one might be interested in domains with infinitely many variables, or domains of variables with an infinite range. Alternatively, one might want to apply objective Bayesianism to full generality of the mathematical framework of §3, or to infinite logical languages (§4). This challenge has been confronted, but at the expense of some objectivity, as we shall now see.

There are two lines of work here, one of which proceeds as follows. Paris and Vencovská (2003) treat problems involving countable logical languages as limiting cases of finite problems. Consider a countably infinite domain $V = \{A_1, A_2, \ldots\}$ of variables taking finitely many values, and schematic evidence $\mathcal{E}$ which may pertain to infinitely many variables. If $V_n = \{A_1, \ldots, A_n\}$ and $\mathcal{E}_n$ is that part of $\mathcal{E}$ that involves only variables in V_n, then $P^{V_n}_{\mathcal{E}_n}(u)$ can be found by maximising entropy as usual (here $u@U \subseteq V_n$). Interestingly—see Paris and Vencovská (2003)—the limit $\lim_{n\to\infty} P^{V_n}_{\mathcal{E}_n}(u)$ exists, so one can define $P^V_{\mathcal{E}}(u)$ to be this limit. Paris and Vencovská (2003) show that this approach can be applied to very simple predicate languages and conjecture that it is applicable more generally to predicate logic.

In the transition from the finite to the infinite, the question arises as to whether countable additivity (introduced in §3) holds. Paris and Vencovská (2003) make no demand that this axiom hold. Indeed, it seems that the type of schematic evidence that they consider cannot be used to express the evidence that an infinite set of outcomes forms a partition. Thus the question of countable additivity cannot be formulated in their framework. In fact, even if one were to extend the framework to formulate the question, the strategy of taking limits would be unlikely to yield probabilities satisfying countable additivity. If the only evidence is that $E_1, \ldots, E_n$ partition the outcome space, maximising entropy will give each event the same probability $1/n$. Taking limits will assign members of an infinite partition probability $\lim_{n\to\infty} 1/n = 0$. But then $\sum_{i=1}^{\infty} P(E_i) = 0 \neq 1$, contradicting countable additivity.

However, not only is countable additivity important from the point of view of mathematical convenience, but according to the standard betting foundations for Bayesian interpretations of probability introduced in §9, countable additivity *must* hold: an agent whose betting quotients are not countably additive can be Dutch booked (Williamson, 1999). Once we accept countable additivity, we are forced either to concede that the strategy of taking limits has only limited applicability, or to reject the method altogether in favour of some alternative, as yet unformulated, strategy. Moreover, as argued in Williamson (1999), we are forced to accept a certain amount of subjectivity: a countably additive distribution of probabilities over a countably infinite partition must award some member of the partition more probability than some other member; but if evidence does not favour any member over any other then it is just a matter of subjective choice as to how one skews the distribution.

The other line of work deals with uncountably infinite domains. Jaynes (1968, §6) presents essentially the following procedure. First find a non-negative real function $P_=(x)$, which we may call the *equivocator* or *invariance* function, that represents the invariances of the problem in question: if $\mathcal{E}$ offers nothing to favour x over y then $P_=(x) = P_=(y)$. Next, find a probability function P satisfying $\mathcal{E}$ that is closest to the invariance function $P_=$, in the sense that it minimises *cross-entropy* distance $d(P, P_=) = \int P(x) \log P(x)/P_=(x)dx$. It is this function that one ought to take as one's belief function $P_\mathcal{E}$.[17]

This approach generalises entropy maximisation on discrete domains. In the case of finite domains $P_=$ can be taken to be the probability function found by maximising entropy subject to no constraints; the probability function $P_\mathcal{E} \in \mathbb{E}$ that is closest to it is just the probability function in $\mathbb{E}$ that has maximum entropy. If the set of variables admits n possible assignments of values, the equivocator $P_=$ can be taken as the function that gives value $1/n$ to each possible assignment v; this is a probability function so $P_\mathcal{E} = P_=$ if there is no evidence whatsoever. In the case of countably infinite domains $P_=$ may not be a probability function: as discussed above $P_=$ must award the same value, k say, to each member of a countable partition; however, such a function cannot be a probability function since countable additivity fails; therefore one must choose a probability function closest to $P_=$. Here we might try to minimise $d(P, P_=) = \sum P(v) \log P(v)/P_=(v) = \sum P(v) \log P(v) - \log k \sum P(v) = \sum P(v) \log P(v) - \log k$; this is minimised just when the entropy $-\sum P(v) \log P(v)$ is maximised. Of course entropy may well be infinite on an infinite partition, so this approach will not work in general; nevertheless a refinement of this kind of approach can yield a procedure for selecting $P_\mathcal{E} \in \mathbb{E}$ that is decisive in many cases (Williamson, 2008a).

By drawing this parallel with the discrete case we can see where problems for objectivity arise in the infinite case: even if the set $\mathbb{E}$ of probability functions compatible with evidence is closed and convex, there may be no probability function in $\mathbb{E}$ closest to $P_=$ or there may be more than one probability function closest to $P_=$. This latter case, non-uniqueness, means subjectivity: the agent can exercise arbitrary choice as which distribution of degrees of belief to select. Subjectivity can also enter at the first stage, choice of $P_=$, since there may be cases in which several different functions represent the invariances of a

[17] Objective Bayesian statisticians have developed a whole host of techniques for obtaining invariance functions and uninformative probability functions—see, e.g., Kass and Wasserman (1996). Berger and Pericchi (2001) discuss the use of such priors in statistics.

problem.[18]

But does such subjectivity really matter? Perhaps not. Although objective Bayesianism often yields objectivity, it can hardly be blamed where little is to be found. If there is nothing to decide between two belief functions, then subjectivity simply does not matter. Under such a view, all the Bayesian positions—strict subjectivism, empirically-based subjective probability and objective Bayesianism—accept the fact that selection of degrees of belief can be a matter of arbitrary choice, they just draw the line in different places as to the extent of subjectivity. Strict subjectivists allow most choice, drawing the line at infringements of the axioms of probability.[19] Proponents of empirically-based subjective probability occupy a half-way house, allowing extensive choice but insisting that evidence of physical probabilities as well as the axioms of probability constrain degrees of belief. Objective Bayesians go furthest by also using logical constraints to narrow down the class of acceptable degrees of belief.

Moreover, arguably the infinite is just a tool to help us reason about the large but finite and discrete universe in which we live (Hilbert, 1925). Just as we create infinite continuous geometries to reason about finite discrete space, we create continuous probability spaces to reason about discrete situations. In which case if subjectivity infects the infinite then we can only conclude that the infinite may not be as effective a tool as we would like for probabilistic reasoning. Such relativity merely urges caution when idealising to the infinite; it does not tell against objective Bayesianism.

§20
Fully Objective Probability

We see then that objectivity is a matter of degree and that while subjectivity may infect some problems, objective Bayesianism yields a high degree of objectivity. We have been focussing on what we might call *epistemic objectivity*, the extent to which an agent's degrees of belief are determined by her evidence. In applications of probability a high degree of epistemic objectivity is an important desideratum: disagreements as to probabilities can be attributed to differences in evidence; by agreeing on evidence consensus can be reached on probabilities.

While epistemic objectivity requires uniqueness relative to evidence, there are stronger grades of objectivity. In particular, the strongest grade of objectivity, *full objectivity*, i.e., uniqueness simpliciter, arouses philosophical interest. Are probabilities uniquely determined, independently of evidence? If two agents disagree as to probabilities must at least one of them be wrong, even if they disagree as to evidence? Intuitively many probabilities are fully objective: there seems to be a fact of the matter as to the probability that an atom of cobalt-60 will decay in 5 years, and there seems to be a fact of the matter as to the chance that a particular roulette wheel will yield a black on the next spin. (A qualification is needed. Chances cannot be quite fully objective inasmuch as they depend on time. There might now be a probability just under 0.5 of cobalt-60

[18]See Gillies (2000, pp. 37–49); Jaynes (1968, §§6–8) and Jaynes (1973). The determination of invariant measures has become an important topic in statistics—see Berger and Pericchi (2001).

[19]Subjectivists usually slip in a few further constraints: e.g., known truths must be given probability 1, and degrees of belief should be updated by Bayesian conditionalisation.

atom decaying in the next five years; after the event, if it has decayed its chance of decaying in that time-frame is 1. Thus chances need to be indexed by time.)

As indicated in §10, objective Bayesianism has the wherewithal to meet the challenge of accounting for intuitions about full objectivity. By considering some ultimate evidence $\hat{\mathcal{E}}$ one can define fully objective probability $\hat{P} = P_{\hat{\mathcal{E}}}$ in terms of the degrees of belief one ought to adopt if one were to have this ultimate evidence. This is the *ultimate belief* notion of chance.

What should be included in $\hat{\mathcal{E}}$? Clearly it should include all information relevant to the domain at time t. To be on the safe side we can take $\hat{\mathcal{E}}$ to include all facts about the universe that are determined by time t—the entire history of the universe up to and including time t. (Remember: this challenge is of philosophical rather than practical interest.)

While the ultimate belief notion of chance is relatively straightforward to state, much needs to be done to show that this type of approach is viable. One needs to show that this notion can capture our intuitions about chance. Moreover, one needs to show that that account is coherent—in particular, one might have concerns about circularity: if probabilistic beliefs are beliefs about probability, yet probability is defined in terms of probabilistic beliefs, then probability appears to be defined in terms of itself.

However, this apparent circularity dissolves when we examine the premisses of this circularity argument more closely. Indeed, at most one premiss can be true. In our framework, 'probability is defined in terms of probabilistic beliefs' is true if we substitute 'fully objective single-case probability' or 'chance' for 'probability' and 'degrees of belief' for 'probabilistic beliefs': chance is defined in terms of degrees of belief. But then the first premiss is false. Degrees of belief are not beliefs about chance, they are partial beliefs about elements of a domain—variables, events or sentences. According to this reading 'probabilistic' modifies 'belief', isolating a type of belief; it does not specify the object of belief. On the other hand, if the first premiss is to be true and 'probabilistic beliefs' are construed as beliefs about probability, then the second premiss is false since chance is not here defined in terms of beliefs about probability. Thus neither reading permits the conclusion that probability is defined in terms of itself.

Note that Bayesian statisticians often consider probability distributions over probability parameters. These *can* be interpreted as degrees of belief about chances, where chances are special degrees of belief. But there is no circularity here either. This is because the degrees of belief about chances are of a higher order than the chances themselves. Consider, for instance, a degree of belief that a particular coin toss will yield heads. The present chance of the coin toss yielding heads can be defined using such degrees of belief. One can then go on to formulate the higher-order degree of belief that the chance of heads is 0.5. But this degree of belief is not used in the (lower order) definition of the chance itself, so there is no circularity. (One can go on to define higher and higher order chances and degrees of belief—regress, rather than circularity, is the obvious problem.)

One can make a stronger case for circularity though. One can read the empirical norm of §13 as saying that degrees of belief ought to be set to chances where they are known (see Williamson, 2005a, §5.3). Under such a reading the concept of rational degree of belief appeals to the notion of chance, yet in this section chances are being construed as special degrees of belief; circularity again. Here circularity is not an artifice of ambiguity of terms like 'probabilistic beliefs'. However, as before, circularity does disappear under closer investigation. One way out is to claim that there are two notions of chance in play: a physical notion which is used in the empirical norm, and an ultimate belief notion which is defined in terms of degrees of belief. But this strategy would not appeal to those who find a physical notion of chance metaphysically or epistemologically dubious. An alternative strategy is to argue that any notion of chance in the formulation of an empirical norm is simply *eliminable*. One can substitute references to chance with references to the *indicators* of chance instead. Intuitively, symmetry considerations, physical laws and observed frequencies all provide some evidence as to chances; one can simply say that an agent's degrees of belief should be appropriately constrained by her evidence of symmetries, laws and frequencies. While this may lead to a rather more complicated formulation of the empirical norm, it is truer to the epistemological route to degrees of belief—the agent has direct evidence of the indicators of chances rather than the chances themselves. Further, it shows how these indicators of chances can actually provide evidence for chances: evidence of frequencies constrains degrees of belief, and chances are just special degrees of belief. Finally, this strategy eliminates circularity, since it shows how degrees of belief can be defined independently of chances. It does, however, pose the challenge of explicating exactly how frequencies, symmetries and so on constrain degrees of belief—a challenge that (as we saw in §18) is not easy to meet.

The ultimate belief notion of chance is not quite fully objective: it is indexed by time. Moreover, if we want a notion of chance defined over infinite domains then, as the arguments of §19 show, subjectivity can creep in, for example in cases—if such cases ever arise—in which the entire history of the universe fails to differentiate between the members of an infinite partition. This mental, ultimate belief notion of chance is arguably more objective than the influential physical notion of chance put forward by David Lewis however (Lewis, 1980, 1994). Lewis accepts a version of the empirical norm which he calls the *Principal Principle*: evidence of chances ought to constrain degrees of belief. However Lewis does not go on to advocate the ultimate belief notion of chance presented here: 'chance is [not] the credence warranted by our total available evidence . . . if our total evidence came from misleadingly unrepresentative samples, that wouldn't affect chance in any way' (Lewis, 1994, p. 475). (Unrepresentative samples do not seem to me to be a real problem for the ultimate belief approach, because the entire history of the universe up to the time in question is likely to contain more information pertinent to an event than simply a small sample frequency—plenty of large samples of relevant events, and plenty of relevant qualitative information, for instance.) Lewis instead takes chances to be products of the best system of laws, the best way of systematising the universe. The problem is that the criteria for comparing systems of laws—a balance between simplicity and strength—seem to be subjective. What counts as simple for a rocket scientist may be complicated for a robot and vice versa.[20] This is not a problem that besets the ultimate belief account: as Lewis accepts, there does seem to be a fact of the matter as to how evidence should inform degrees of belief. Thus an ultimate belief notion of chance, despite being a mental rather than physical notion, suffers less from subjectivity than Lewis' theory.

[20]In response Lewis (1994, p. 479) just plays the optimism card: 'if nature is kind to us, the problem needn't arise.'

Note that Lewis' approach also suffers from a type of circularity known as *undermining*. Because chances for Lewis are analysed in terms of laws, they depend not only on the past and present state of the universe, but also on the future of the universe: 'present chances are given by probabilistic laws, plus present conditions to which those laws are applicable, and ... those laws obtain in virtue of the fit of candidate systems to the whole of history' (Lewis, 1994, p. 482). Of course, non-actual futures (i.e., series of events which differ from the way in which the universe will actually turn out) must have positive chance now, for otherwise the notion of chance would be redundant. Thus there is now a positive chance of events turning out in the future in such a way that present chances turn out differently. But this yields a paradox: present chances cannot turn out differently to what they actually are. Lewis (1994) has to modify the Principal Principle to avoid a formal contradiction, but this move does not resolve the intuitive paradox. In contrast, under the ultimate belief account present chances depend on just the past and the present state of the universe, not the future, so present chances cannot undermine themselves.

§21
Probability Logic

There are increasing demands from researchers in artificial intelligence for formalisms for normative reasoning that combine probability and logic. Purely probabilistic techniques work quite well in many areas but fail to exploit logical relationships that obtain in particular problems. Thus, for example, probabilistic techniques are applied widely in natural language processing (Manning and Schütze, 1999), with some success, yet largely without exploiting logical sentence structure. On the other hand, purely logical techniques take problem structure into account without being able to handle the many uncertainties inherent in practical problem solving. Thus automated proof systems for mathematical reasoning (Quaife, 1992; Schumann, 2001) depend heavily on implementing logics but often fail to prioritise searches that are most likely to be successful. It is natural to suppose that systems which combine probability and logic will yield improved results. Formalisms that combine probability and logic would also be applicable to many new problems in bioinformatics (Durbin et al., 1999), from inducing protein folding from noisy relational data to forecasting toxicity from uncertain evidence of deterministic chemical reactions in cell metabolism.

In a *probability logic*, or *progic* for short, probability is combined with logic in one or more of the following two ways:

External: probabilities are *attached to* sentences of a logical language,

Internal: sentences *incorporate* statements about probabilities.

In an *external* progic, entailment relationships take the form:

$$\varphi_1{}^{X_1}, \ldots, \varphi_n{}^{X_n} \approx \psi^Y$$

Here $\varphi_1, \ldots, \varphi_n, \psi \in S\mathcal{L}$ are sentences of a logical language $\mathcal{L}$ which does not contain probabilities and $X_1, \ldots, X_n, Y \in [0, 1]$ are sets of probabilities. For example if $\mathcal{L} = \{A_1, A_2, A_3, A_4, A_5\}$ is a propositional language on propositional variables $A_1, \ldots, A_5$, we might be interested in what set Y of probabilities to attach to the conclusion in

$$A_1 \wedge \neg A_2{}^{.9}, (\neg A_4 \vee A_3) \rightarrow A_2{}^{.2}, A_5 \vee A_3{}^{.3}, A_4{}^{.7} \approx\!\!\!\!\models A_5 \rightarrow A_1{}^{Y}.$$

In an *internal* progic, entailment relationships take the form:

$$\varphi_1, \ldots, \varphi_n \approx\!\!\!\!\models \psi$$

where $\varphi_1, \ldots, \varphi_n, \psi \in S\mathcal{L}_P$ are sentences of a logical language $\mathcal{L}_P$ which contains probabilities. $\mathcal{L}_P$ might be a first-order language with equality containing a (probability) function P, predicates U_1, U_2, U_3 and constants sorted into individuals t_i, events e_i and real numbers $x_i \in [0,1]$, and we might want to know whether

$$P(e_1) = x_1 \vee U_1(t_3), \neg P(e_2) = x_1 \rightarrow W(t_5) \approx\!\!\!\!\models U_1(t_5).$$

Note that an internal progic might have several probability functions, each with a different interpretation.

In a *mixed* progic, the probabilities may appear both internally and externally. An entailment relationship takes the form

$$\varphi_1^{X_1}, \ldots, \varphi_n{}^{X_n} \approx\!\!\!\!\models \psi^Y$$

where $\varphi_1, \ldots, \varphi_n, \psi \in S\mathcal{L}_P$ are sentences of a logical language $\mathcal{L}_P$ which contains probabilities.

There are two main questions to be dealt with when providing semantics for a progic: how are the probabilities to be interpreted? what is the meaning of the entailment relation symbol $\approx\!\!\!\!\models$?

The *standard probabilistic semantics* remains neutral about the interpretation of the probabilities and deals with entailment thus:

External: $\varphi_1^{X_1}, \ldots, \varphi_n{}^{X_n} \approx\!\!\!\!\models \psi^Y$ holds if and only if every probability function P that satisfies the left-hand side (i.e., $P(\varphi_1) \in X_1, \ldots, P(\varphi_n) \in X_n$) also satisfies the right-hand side (i.e., $P(\psi) \in Y$).

Internal: $\varphi_1, \ldots, \varphi_n \approx\!\!\!\!\models \psi$ if and only if every $\mathcal{L}_P$-model of the left-hand side in which P is interpreted as a probability function is also a model of the right-hand side.

The difficulty with the standard semantics for an external progic is that of *underdetermination*. Given some premiss sentences $\varphi_1, \ldots, \varphi_n$ and their probabilities $X_1, \ldots, X_n$ we often want to know what *single* probability y to give to a conclusion sentence ψ of interest. However, the standard semantics may give no answer to this question: often $\varphi_1^{X_1}, \ldots, \varphi_n{}^{X_n} \approx\!\!\!\!\models \psi^Y$ for a nonsingleton $Y \subseteq [0,1]$, because probability functions that satisfy the left-hand side disagree as to the probability they award to ψ on the right-hand side. The premisses underdetermine the conclusion. Consequently an alternative semantics is often preferred.

According to the *objective Bayesian semantics* for an external progic on a finite propositional language $L = \{A_1, \ldots, A_N\}$, $\varphi_1^{X_1}, \ldots, \varphi_n{}^{X_n} \approx\!\!\!\!\models \psi^Y$ if and only if an agent whose evidence is summed up by the constraints on the left-hand side (so who ought to believe φ_1 to degree in $X_1, \ldots, \varphi_n$ to degree in X_n) ought to believe ψ to degree in Y. As long as the constraints $\varphi_1^{X_1}, \ldots, \varphi_n{}^{X_n}$ are consistent, there will be a unique function P that maximises entropy and a unique $y \in [0,1]$ such that $P(\psi) = y$, so there is no problem of underdetermination.

I shall briefly sketch just three of the principal proposals in this area.[21]

Colin Howson put forward his account of the relationship between probability and logic in Howson (2001); Howson (2003) and Howson (2008). Howson interprets probability as follows: 'the agent's probability is the odds, or the betting quotient, they currently believe fair, with the sense of 'fair' that there is no calculable advantage to either side of a bet at those odds' (Howson, 2001, 143). The connection with logic is forged by introducing the concept of consistency of betting quotients: a set of betting quotients is consistent if it can be extended to a single-valued function on all the propositions of a given logical language $\mathcal{L}$ which satisfies certain regularity properties. Howson then shows that an assignment of betting quotients is consistent if and only if it is satisfiable by a probability function (Howson, 2001, Theorem 1). Having developed a notion of consistency, Howson shows that this leads naturally to an external progic with the standard semantics: consequence is defined in terms of satisfiability by probability functions, as outlined above (Howson, 2001, 150).

In Halpern (2003), Joseph Halpern studies the standard semantics for internal progics. In the propositional case, $\mathcal{L}$ is a propositional language extended by permitting linear combinations of probabilities $\sum_{i=1}^n a_i P_i(\psi_i) > b$, where $a_1, \ldots, a_n, b \in \mathbb{R}$ and $P_1, \ldots, P_n$ are probability functions each of which represents the degrees of belief of an agent and which are defined over sentences ψ of $\mathcal{L}$ (Halpern, 2003, §7.3). This language allows nesting of probabilities: for example $P_1(\neg(P_2(\varphi) > 1/3)) > 1/2$ represents 'with degree more than a half, agent 1 believes that agent 2's degree of belief in φ is less than or equal to $\frac{1}{3}$.' Note, though, that the language cannot represent probabilistic independencies, which are expressed using multiplication rather than linear combination of probabilities, such as $P_1(\varphi \wedge \psi) = P_1(\varphi)P_1(\psi)$. Halpern provides a possible-worlds semantics for the resulting logic: given a space of possible worlds, a probability measure $\mu_{w,i}$ over this space for each possible world and agent, and a valuation function π_w for each possible world, $P_1(\psi) > 1/2$ is true at a world w if the measure $\mu_{w,1}$ of the set of possible words at which ψ is true is greater than half, $\mu_{w,1}(\{w' : \pi_{w'}(\psi) = 1\}) > 1/2$. Consequence is defined straightforwardly in terms of satisfiability by worlds.

Halpern later extends the above propositional language to a first-order language and introduces frequency terms $\|\psi\|_X$, interpreted as 'the frequency with which ψ holds when variables in X are repeatedly selected at random' (Halpern, 2003, §10.3). Linear combinations of frequencies are permitted, as well as linear combinations of degrees of belief. When providing the semantics for this language, one must provide an interpretation for frequency terms, a probability measure over the domain of the language.

In Paris (1994), Jeff Paris discusses external progics in detail, in conjunction with the objective Bayesian semantics. In the propositional case, Paris proposes a number of common sense desiderata which ought to be satisfied by any method for picking out a most rational belief function for the objective Bayesian semantics, and goes on to show that the Maximum Entropy Principle is the *only* method that satisfies these desiderata (Paris, 1994, Theorem 7.9; Paris and

[21] Williamson (2002) presents a more comprehensive survey.

Vencovská, 2001). Later Paris shows how an external progic can be defined over the sentences of a first order logic—such a function is determined by its values over quantifier-free sentences (Paris, 1994, Chapter 11; Gaifman, 1964). Paris then introduces the problem of learning from experience: what value should an agent give to $P(U(t_{n+1})|\pm U(t_1) \wedge \cdots \wedge \pm U(t_n))$, that is, to what extent should she believe a new instance of U, given n observed instances (Paris, 1994, Chapter 12)? As mentioned in §§18, 19, Paris and Vencovská (2003) and Williamson (2008a) suggest that the Maximum Entropy Principle may be extended to the first-order case to address this problem, though by appealing to rather different strategies.

In the case of the standard semantics one might look for a traditional proof theory to accompany the semantics:

External: Given $\varphi_1, \ldots \varphi_n \in S\mathcal{L}, x_1, \ldots, x_n \in [0, 1]$, find a mechanism for generating all ψ^Y such that $\varphi_1^{X_1}, \ldots, \varphi_n{}^{X_n} \approx \psi^Y$.

Internal: Given $\varphi_1, \ldots \varphi_n \in S\mathcal{L}_P$, find a mechanism for generating all $\psi \in S\mathcal{L}_P$ such that $\varphi_1, \ldots, \varphi_n \approx \psi$.

In a sense this is straightforward: the premisses imply the conclusion just if the conclusion follows from the premisses and the axioms of probability by deductive logic. Fagin et al. (1990) produced a traditional proof theory for the standard probabilistic semantics, for an internal propositional progic. As with propositional logic, deciding satisfiability is NP-complete. Halpern (1990) discusses a progic which allows reasoning about both degrees of belief and frequencies. In general, no complete axiomatisation is possible, though axiom systems are provided in special cases where complete axiomatisation is possible. Abadi and Halpern (1994) consider first-order degree of belief and frequency logics separately, and show that they are highly undecidable. Halpern (2003) presents a general overview of this line of work.

Paris and Vencovská (1990) made a start at a traditional proof theory for a type of objective Bayesian progic, but express some scepticism as to whether the goal of a traditional proof system can be achieved.

A traditional proof theory, though interesting, is often not what is required in applications of an external progic. To reiterate, given some premiss sentences $\varphi_1, \ldots, \varphi_n$ and sets of probabilities $X_1, \ldots, X_n$ we often want to know what set of probabilities Y to give to a conclusion sentence ψ of interest—not to churn out all ψ^Y that follow from the premisses. Objective Bayesianism provides semantics for this problem, and it is an important question as to whether there is a calculus that accompanies this semantics:

Obprogic: Given $\varphi_1, \ldots, \varphi_n, X_1, \ldots, X_n, \psi$, find an appropriate Y such that $\varphi_1^{X_1}, \ldots, \varphi_n{}^{X_n} \approx \psi^Y$.

By 'appropriate Y' here we mean the narrowest such Y: the entailment trivially holds for $Y = [0, 1]$; a maximally specific Y will be of more interest.

It is known that even finding an approximate solution to this problem is NP-complete (Paris, 1994, Theorem 10.6). Hence the best one can do is to find an algorithm that is scalable in a range of natural problems, rather than tractable in every case. The approach of Williamson (2005a) deals with the propositional case but does not take the form of a traditional logical proof theory, involving axioms and rules of inference. Instead, the proposal is to apply the computational methods of §17 to find an objective Bayesian net—a Bayesian net representation of the P that satisfies constraints $P(\varphi_1) \in X_1, \ldots, P(\varphi_n) \in X_n$ and maximises entropy—and then to use this net to calculate $P(\psi)$. The advantage of using Bayesian nets is that, if sufficiently sparse, they allow the efficient representation of a probability function and efficient methods for calculating marginal probabilities of that function. In this context, the net is sparse and the method scalable in cases where each sentence involves few propositional variables in comparison with the size of the language.

Consider an example. Suppose we have a propositional language $\mathcal{L} = \{A_1, A_2, A_3, A_4, A_5\}$ and we want to find Y such that

$$A_1 \wedge \neg A_2,^{.9} (\neg A_4 \vee A_3) \rightarrow A_2,^{.2} A_5 \vee A_3,^{.3} A_4{}^{.7} \approx\!\!\!\!| \; A_5 \rightarrow A_1.^Y$$

According to our semantics we must find P that maximises

$$H = -\sum P(\pm A_1 \wedge \pm A_2 \wedge \pm A_3 \wedge \pm A_4 \wedge \pm A_5) \log P(\pm A_1 \wedge \pm A_2 \wedge \pm A_3 \wedge \pm A_4 \wedge \pm A_5)$$

subject to the constraints,

$$P(A_1 \wedge \neg A_2) = .9, P((\neg A_4 \vee A_3) \rightarrow A_2) = .2, P(A_5 \vee A_3) = .3, P(A_4) = .7$$

One could find P by directly using numerical optimisation techniques or Lagrange multiplier methods. However, this approach would not be feasible on large languages—already we would need to optimise with respect to 2^5 parameters $P(\pm A_1 \wedge \pm A_2 \wedge \pm A_3 \wedge \pm A_4 \wedge \pm A_5)$.

Instead take the approach of §17:

Step 1: Construct an undirected *constraint graph*, Fig. 1, by linking variables that occur in the same constraint.

As mentioned, the constraint graph satisfies a key property, namely, separation in the constraint graph implies conditional independence for the entropy maximising probability function P. Thus A_2 separates A_5 from A_1, so $A_1 \perp\!\!\!\perp A_5 \mid A_2$, ($P$ renders A_1 probabilistically independent of A_5 conditional on A_2).

Step 2: Transform this into a *directed constraint graph*, Fig. 2.

Now D-separation, a directed version of separation (Pearl, 1988, §3.3), implies conditional independence for P. Having found a directed acyclic graph which satisfies this property we can construct a Bayesian net by augmenting the graph with conditional probability distributions:

Step 3: Form a Bayesian network by determining parameters $P(A_i|par_i)$ that maximise entropy.

Here the par_i are the states of the parents of A_i. Thus we need to determine $P(A_1), P(A_2|\pm A_1), P(A_3|\pm A_2), P(A_4|\pm A_3 \wedge \pm A_2), P(A_5|\pm A_3)$. This can be done by reparameterising the entropy equation in terms of these conditional probabilities and then using Lagrange multiplier methods or numerical optimisation techniques. This representation of P will be efficient if the graph is sparse, that is, if each constraint sentence φ_i involves few propositional variables in comparison with the size of the language.

STEP 4: Simplify ψ into a disjunction of mutually exclusive conjunctions $\bigvee \sigma_j$ (e.g., full disjunctive normal form) and calculate $P(\psi) = \sum P(\sigma_j)$ by using standard Bayesian net algorithms to determine the marginals $P(\sigma_j)$.

In our example,

$$\begin{aligned} P(A_5 \rightarrow A_1) &= P(\neg A_5 \vee A_1) \\ &= P(\neg A_5 \wedge A_1) + P(A_5 \wedge A_1) + P(\neg A_5 \wedge \neg A_1) \\ &= P(\neg A_5|A_1)P(A_1) + P(A_5|A_1)P(A_1) + P(\neg A_5|\neg A_1)P(\neg A_1) \\ &= P(A_1) + P(\neg A_5|\neg A_1)(1 - P(A_1)). \end{aligned}$$

We thus require only two Bayesian net calculations to determine $P(A_1)$ and $P(\neg A_5|\neg A_1)$. These calculations can be performed efficiently if the graph is sparse and ψ involves few propositional variables relative to the size of the domain.

A major challenge for the objective Bayesian approach is to see whether potentially efficient procedures can be developed for first-order predicate logic. Williamson (2008a) takes a step in this direction by showing that objective Bayesian nets, and a generalisation, *objective credal nets*, can in principle be applied to first-order predicate languages.

PART IV
IMPLICATIONS FOR THE PHILOSOPHY OF MATHEMATICS

Probability theory is a part of mathematics; it should be uncontroversial then that the philosophy of probability is relevant to the philosophy of mathematics. Unfortunately, though, philosophers of mathematics tend to pass over the philosophy of probability, viewing it as a branch of the philosophy of science rather than the philosophy of mathematics. Here I shall attempt to redress the balance by suggesting ways in which the philosophy of probability can suggest new directions to the philosophy of mathematics in general.

§22
THE ROLE OF INTERPRETATION

One potential interaction concerns the existence of mathematical entities. Philosophers of probability tackle the question of the existence of probabilities within the context of an interpretation. Questions like 'what are probabilities?' and 'where are they?' receive different answers according to the interpretation of probability under consideration. There is little dispute that axioms of probability admit of more than one interpretation: Bayesians argue convincingly that rational degrees of belief satisfy the axioms of probability; frequentists argue convincingly that limiting relative frequencies satisfy the axioms (except the axiom of countable additivity). The debate is not so much about finding *the* interpretation of probability, but about which interpretation is *best* for particular applications of probability—applications as diverse as those in statistics, number theory, machine learning, epistemology and the philosophy of science.

Now according to the Bayesian interpretation probabilities are mental entities, according to frequency theories they are features of collections of physical outcomes, and according to propensity theories they are features of physical experimental set-ups or of single-case events. So we see that an interpretation is required before one can answer questions about existence. The uninterpreted mathematics of probability is treated in an *if-then*-ist way: if the axioms hold then Bayes' theorem holds; degrees of rational belief satisfy the axioms; therefore degrees of rational belief satisfy Bayes' theorem.

The question thus arises as to whether it may *in general* be most productive to ask what mathematical entities are *within the context of an interpretation*. It may make more sense to ask 'what kind of thing is a Hilbert space in the epistemic interpretation of quantum mechanics?' than 'what kind of thing is a Hilbert space?' In mathematics it is crucial to ask questions at the right level of generality; so too in the philosophy of mathematics.

Such a shift in focus from abstraction towards interpretation introduces important challenges. For example, the act of interpretation is rarely a straightforward matter—it typically requires some sort of idealisation. While elegance plays a leading role in the selection of mathematics, the world is rather more messy, and any mapping between the two needs a certain leeway. Thus rational degrees of belief are idealised as real numbers, even though an agent would be irrational to worry about the $10^{10^{10}}$-th decimal place of her degree of belief; frequencies are construed as limits of finite relative frequencies, even though that limit is never actually reached. When assessing an interpretation, the suitability of its associated idealisations are of paramount importance. If it makes a *substantial* difference what the $10^{10^{10}}$-th decimal place of a degree of belief is, then so much the worse for the Bayesian interpretation of probability. Similarly when interpreting arithmetic or set theory: if it matters that a large collection of objects is not in fact denumerable then one should not treat it as the domain of an interpretation of Peano arithmetic; if it matters that the collection is not in fact an object distinct from its members then one should not treat it as a set. A first challenge, then, is to elucidate the role of idealisation in interpretations.

A second challenge is to demarcate the interpretations that imbue existence on mathematical entities from those that don't. While some interpretations construe mathematical entities as worldly things, some construe mathematical entities in terms of other uninterpreted mathematical entities. To take a simple example, one may appeal to affine transformations to interpret the axioms of group theory. In order to construe this group as existing, one must go on to say something about the existence of the transformations: one needs a chain of interpretations that is grounded in worldly things. In the absence of such grounding, the interpretation fails to impart existence. These interpretations within mathematics are rather different from the interpretations that are grounded in our messy world, in that they tend not to involve idealisation: the transformations really do form a group. But of course the line between world and mathematics can be rather blurry, especially in disciplines like theoretical physics: are quantum fields part of the world, or do they require further interpretation?[22]

This shift in focus from abstraction to interpretation is ontological, but not epistemological. That mathematical entities must be interpreted to exist does not mean that uninterpreted mathematics does not qualify as knowledge. Tak-

[22] Corfield (2003, Part IV) discusses interpretations within mathematics.

ing an *if-then*-ist view of uninterpreted mathematics, knowledge is accrued if one knows that the consequent does indeed follow from the antecedent, and the role of proof is of course crucial here.[23]

§23
The Epistemic view of Mathematics

But there is undoubtedly more to mathematics than a collection of *if-then* statements and a further analogy with Bayesianism suggests a more sophisticated philosophy. Under the Bayesian view probabilities are rational degrees of belief, a feature of an agent's epistemic state; they do not exist independently of agents. According to objective Bayesianism probabilities are also objective, in the sense that two agents with the same background information have little or no room for disagreement as to the probabilities. This objectivity is a result of the fact that an agent's degrees of belief are heavily constrained by the extent and limitations of her empirical evidence.

Perhaps mathematics is also purely epistemic, yet objective. Just as Bayesianism considers probabilistic beliefs to be a type of belief—point-valued degrees of belief—rather than beliefs about agent-independent probabilities, mathematical beliefs may also be a type of belief, rather than beliefs about uninterpreted mathematical entities. Just as probabilistic beliefs are heavily constrained, so too mathematical beliefs are heavily constrained. Perhaps so heavily constrained that mathematics turns out to be fully objective, or nearly fully objective (there may be room for subjective disagreement about some principles, such as the continuum hypothesis).[24]

The constraints on mathematical beliefs are the bread and butter of mathematics. Foremost, of course, mathematical beliefs need to be useful. They need to generate good predictions and explanations, both when applied to the real world, i.e., to interpreted mathematical entities, and when applied within mathematics itself. The word 'good' itself encapsulates several constraints: predictions and explanations must achieve a balance of being accurate, interesting, powerful, simple and fruitful, and must be justifiable using two modes of reasoning: proof and interpretation. Finally sociological constraints may have some bearing (e.g. mathematical beliefs need to further mathematicians in their careers and power struggles; the development of mathematics is no doubt constrained by the fact that the most popular conferences are in beach locations)—the question is how big a role such constraints play.

The objective Bayesian analogy then leads to an *epistemic* view of mathematics characterised by the following hypotheses:[25]

Convenience: Mathematical beliefs are convenient, because they admit good explanations and predictions within mathematics itself and also within its grounding interpretations.

[23]See Awodey (2004) for a defence of a type of *if-then*-ism.

[24]Paseau (2005) emphasises the interpretation of mathematics. In his terminology, I would be suggesting a reinterpretation of mathematics in terms of rational beliefs. This notion of *re*interpretation requires there to be some natural or default interpretation that is to be superseded. But as Paseau (2005, pp. 379–380) himself notes, it is by no means clear that there is such a default interpretation.

[25]An analogous epistemic view of causality is developed in Williamson (2005a, Chapter 9).

Explanation: We have mathematical beliefs because of this convenience, not because uninterpreted mathematical entities correspond to physical things that we experience, nor because such entities correspond to platonic things that we somehow intuit.

Objectivity: The strength of the constraints on mathematical beliefs renders mathematics an objective, or nearly objective, activity.

Under the epistemic view, then, mathematics is like an axe. It is a tool whose design is largely determined by constraints placed on it.[26] Just as the design of an axe is roughly determined by its use (chopping wood) and demands on its strength and longevity, so too mathematics is roughly determined by its use (prediction and explanation) and high standard of certainty as to its conclusions. No wonder that mathematicians working independently end up designing similar tools.

§24
Conclusion

If probability is to be applied it must be interpreted. Typically we are interested in single-case probabilities—e.g., the probability that I will live to the age of 80, the probability that my car will break down today, the probability that quantum mechanics is true. The Bayesian interpretation tells us what such probabilities are: they are rational degrees of belief.

Subjective Bayesianism has the advantage that it is easy to justify—the Dutch book argument is all that is needed. But subjective Bayesianism does not successfully capture our intuition that many probabilities are objective.

If we move to objective Bayesianism what we gain in terms of objectivity, we pay for in terms of hard graft to address the challenges outlined in Part III. (For this reason, many Bayesians are subjectivist in principle but tacitly objectivist in practice.) These are just challenges though; none seem to present insurmountable problems. They map out an interesting and important research programme rather than reasons to abandon any hope of objectivity.

The two principal ideas of this chapter—that of interpretation and that of objectively-determined belief—are key if we are to understand probability. I have suggested that they might also offer some insight into mathematics in general.

Acknowledgements

I am very grateful to Oxford University Press for permission to reprint material from Williamson (2005a) in Part I and Part II of this chapter, and for permission to reprint material from Williamson (2006) in Part IV. I am also grateful to the Leverhulme Trust for a research fellowship supporting this research.

[26] Marquis (1997, p. 252) discusses the claim that mathematics contains tools or instruments as well as an independent reality of uninterpreted mathematical entities. The epistemic position, however, is purely instrumentalist: there are tools but no independent reality. As Marquis notes, the former view has to somehow demarcate between mathematical objects and tools—by no means an easy task.

References

Abadi, M. and Halpern, J. Y. (1994). Decidability and expressiveness for first-order logics of probability. *Information and Computation*, 112(1):1–36.

Awodey, S. (2004). An answer to Hellman's question: 'Does category theory provide a framework for mathematical structuralism?'. *Philosophia Mathematica (3)*, 12:54–64.

Berger, J. O. and Pericchi, L. R. (2001). Objective Bayesian methods for model selection: introduction and comparison. In Lahiri, P., editor, *Model Selection*, volume 38 of *Monograph Series*, pages 135–207. Beachwood, Ohio. Institute of Mathematical Statistics Lecture Notes.

Bernoulli, J. (1713). *Ars Conjectandi*. The Johns Hopkins University Press, Baltimore, 2006 edition. Trans. Edith Dudley Sylla.

Billingsley, P. (1979). *Probability and measure*. John Wiley and Sons, New York, third (1995) edition.

Carnap, R. (1952). *The continuum of inductive methods*. University of Chicago Press, Chicago IL.

Carnap, R. (1971). A basic system of inductive logic part 1. In Carnap, R. and Jeffrey, R. C., editors, *Studies in inductive logic and probability*, volume 1, pages 33–165. University of California Press, Berkeley CA.

Church, A. (1936). An unsolvable problem of elementary number theory. *American Journal of Mathematics*, 58:345–363.

Corfield, D. (2003). *Towards a philosophy of real mathematics*. Cambridge University Press, Cambridge.

de Finetti, B. (1937). Foresight. its logical laws, its subjective sources. In Kyburg, H. E. and Smokler, H. E., editors, *Studies in subjective probability*, pages 53–118. Robert E. Krieger Publishing Company, Huntington, New York, second (1980) edition.

Della Pietra, S., Della Pietra, V. J., and Lafferty, J. D. (1997). Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):380–393.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1999). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge.

Earman, J. (1992). *Bayes or bust?* MIT Press, Cambridge MA.

Fagin, R., Halpern, J. Y., and Megiddo, N. (1990). A logic for reasoning about probabilities. *Information and Computation*, 87(1-2):277–291.

Fetzer, J. H. (1982). Probabilistic explanations. *Philosophy of Science Association*, 2:194–207.

Gaifman, H. (1964). Concerning measures in first order calculi. *Israel Journal of Mathematics*, 2:1–18.

Gaifman, H. and Snir, M. (1982). Probabilities over rich languages. *Journal of Symbolic Logic*, 47(3):495–548.

Gillies, D. (2000). *Philosophical theories of probability*. Routledge, London and New York.

Hacking, I. (1975). *The emergence of probability*. Cambridge University Press, Cambridge.

Halpern, J. Y. (1990). An analysis of first-order logics of probability. *Artificial Intelligence*, 46:311–350.

Halpern, J. Y. (2003). *Reasoning about uncertainty*. MIT Press, Cambridge

MA.
Halpern, J. Y. and Koller, D. (1995). Representation dependence in probabilistic inference. In Mellish, C. S., editor, *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI 95)*, pages 1853–1860. Morgan Kaufmann, San Francisco CA.
Hilbert, D. (1925). On the infinite. In Benacerraf, P. and Putnam, H., editors, *Philosophy of mathematics: selected readings*. Cambridge University Press (1983), Cambridge, second edition.
Howson, C. (2001). The logic of Bayesian probability. In Corfield, D. and Williamson, J., editors, *Foundations of Bayesianism*, pages 137–159. Kluwer, Dordrecht.
Howson, C. (2003). Probability and logic. *Journal of Applied Logic*, 1(3-4):151–165.
Howson, C. (2008). Can logic be combined with probability? probably. *Journal of Applied Logic*, doi:10.1016/j.jal.2007.11.003.
Howson, C. and Urbach, P. (1989). *Scientific reasoning: the Bayesian approach*. Open Court, Chicago IL, second (1993) edition.
Hunter, D. (1989). Causality and maximum entropy updating. *International Journal in Approximate Reasoning*, 3:87–114.
Jaynes, E. T. (1957). Information theory and statistical mechanics. *The Physical Review*, 106(4):620–630.
Jaynes, E. T. (1968). Prior probabilities. *IEEE Transactions Systems Science and Cybernetics*, SSC-4(3):227.
Jaynes, E. T. (1973). The well-posed problem. *Foundations of Physics*, 3:477–492.
Jaynes, E. T. (1979). Where do we stand on maximum entropy? In Levine, R. and Tribus, M., editors, *The maximum entropy formalism*, page 15. MIT Press, Cambridge MA.
Jaynes, E. T. (1988). The relation of Bayesian and maximum entropy methods. In Erickson, G. J. and Smith, C. R., editors, *Maximum-entropy and Bayesian methods in science and engineering*, volume 1, pages 25–29. Kluwer, Dordrecht.
Jaynes, E. T. (2003). *Probability theory: the logic of science*. Cambridge University Press, Cambridge.
Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91:1343–1370.
Keynes, J. M. (1921). *A treatise on probability*. Macmillan (1948), London.
Kolmogorov, A. N. (1933). *The foundations of the theory of probability*. Chelsea Publishing Company (1950), New York.
Laplace (1814). *A philosophical essay on probabilities*. Dover (1951), New York. Pierre Simon, marquis de Laplace.
Lewis, D. K. (1980). A subjectivist's guide to objective chance. In *Philosophical papers*, volume 2, pages 83–132. Oxford University Press (1986), Oxford.
Lewis, D. K. (1994). Humean supervenience debugged. *Mind*, 412:471–490.
Manning, C. D. and Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press, Cambridge MA.
Marquis, J.-P. (1997). Abstract mathematical tools and machines for mathematics. *Philosophia Mathematica (3)*, 5:250–272.
Miller, D. (1994). *Critical rationalism: a restatement and defence*. Open Court, Chicago IL.

Nagl, S., Williams, M., and Williamson, J. (2008). Objective Bayesian nets for systems modelling and prognosis in breast cancer. In Holmes, D. and Jain, L., editors, *Innovations in Bayesian networks: theory and applications*. Springer.

Paris, J. B. (1994). *The uncertain reasoner's companion*. Cambridge University Press, Cambridge.

Paris, J. B. and Vencovská, A. (1990). A note on the inevitability of maximum entropy. *International Journal of Approximate Reasoning*, 4:181–223.

Paris, J. B. and Vencovská, A. (1997). In defence of the maximum entropy inference process. *International Journal of Approximate Reasoning*, 17:77–103.

Paris, J. B. and Vencovská, A. (2001). Common sense and stochastic independence. In Corfield, D. and Williamson, J., editors, *Foundations of Bayesianism*, pages 203–240. Kluwer, Dordrecht.

Paris, J. B. and Vencovská, A. (2003). The emergence of reasons conjecture. *Journal of Applied Logic*, 1(3–4):167–195.

Paseau, A. (2005). Naturalism in mathematics and the authority of philosophy. *British Journal for the Philosophy of Science*, 56:377–396.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, San Mateo CA.

Popper, K. R. (1934). *The Logic of Scientific Discovery*. Routledge (1999), London. With new appendices of 1959.

Popper, K. R. (1959). The propensity interpretation of probability. *British Journal for the Philosophy of Science*, 10:25–42.

Popper, K. R. (1983). *Realism and the aim of science*. Hutchinson, London.

Popper, K. R. (1990). *A world of propensities*. Thoemmes, Bristol.

Quaife, A. (1992). *Automated development of fundamental mathematical theories*. Kluwer, Dordrecht.

Ramsey, F. P. (1926). Truth and probability. In Kyburg, H. E. and Smokler, H. E., editors, *Studies in subjective probability*, pages 23–52. Robert E. Krieger Publishing Company, Huntington, New York, second (1980) edition.

Reichenbach, H. (1935). *The theory of probability: an inquiry into the logical and mathematical foundations of the calculus of probability*. University of California Press (1949), Berkeley and Los Angeles. Trans. Ernest H. Hutten and Maria Reichenbach.

Rosenkrantz, R. D. (1977). *Inference, method and decision: towards a Bayesian philosophy of science*. Reidel, Dordrecht.

Schumann, J. M. (2001). *Automated theorem proving in software engineering*. Springer-Verlag.

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423 and 623–656.

Venn, J. (1866). *Logic of chance: an essay on the foundations and province of the theory of probability*. Macmillan, London.

von Mises, R. (1928). *Probability, statistics and truth*. Allen and Unwin, London, second (1957) edition.

von Mises, R. (1964). *Mathematical theory of probability and statistics*. Academic Press, New York.

Williamson, J. (1999). Countable additivity and subjective probability. *British Journal for the Philosophy of Science*, 50(3):401–416.

Williamson, J. (2002). Probability logic. In Gabbay, D., Johnson, R., Ohlbach, H. J., and Woods, J., editors, *Handbook of the logic of argument and inference:*

the turn toward the practical, pages 397–424. Elsevier, Amsterdam.

Williamson, J. (2005a). *Bayesian nets and causality: philosophical and computational foundations*. Oxford University Press, Oxford.

Williamson, J. (2005b). Objective Bayesian nets. In Artemov, S., Barringer, H., d'Avila Garcez, A. S., Lamb, L. C., and Woods, J., editors, *We Will Show Them! Essays in Honour of Dov Gabbay*, volume 2, pages 713–730. College Publications, London.

Williamson, J. (2006). From Bayesianism to the epistemic view of mathematics. *Philosophia Mathematica (III)*, 14(3):365–369.

Williamson, J. (2007a). Inductive influence. *British Journal for the Philosophy of Science*, 58(4):689–708.

Williamson, J. (2007b). Motivating objective Bayesianism: from empirical constraints to objective probabilities. In Harper, W. L. and Wheeler, G. R., editors, *Probability and Inference: Essays in Honour of Henry E. Kyburg Jr.*, pages 151–179. College Publications, London.

Williamson, J. (2008a). Objective Bayesian probabilistic logic. *Journal of Algorithms in Cognition, Informatics and Logic*, in press.

Williamson, J. (2008b). Objective Bayesianism, Bayesian conditionalisation and voluntarism. *Synthese*, in press.

Williamson, J. (2008c). Objective Bayesianism with predicate languages. *Synthese*, 163(3):341–356.