Entropy and Insufficient Reason: A Note on the Judy Benjamin Problem

Anubav Vasudevan

Abstract

One well-known objection to the principle of maximum entropy is the so-called Judy Benjamin problem, first introduced by van Fraassen ([1981]). The problem turns on the apparently puzzling fact that, on the basis of information relating an event's conditional probability, the maximum entropy distribution will almost always assign to the event conditionalized on a probability strictly less than that assigned to it by the uniform distribution. In this paper, I present an analysis of the Judy Benjamin problem that can help to make sense of this seemingly odd feature of maximum entropy inference. My analysis is based on the claim that, in applying the principle of maximum entropy, Judy Benjamin is not acting out of a concern to maximize uncertainty in the face of new evidence, but is rather exercising a certain brand of epistemic charity towards her informant. This epistemic charity takes the form of an assumption on the part of Judy Benjamin that her informant's evidential report leaves out no relevant information. Such a reconceptualization of the motives underlying Judy Benjamin's appeal to the principle of maximum entropy can help to further our understanding of the true epistemological grounds of this principle and, in particular, can shed light on the nature of the relationship between the principle of maximum entropy and the Laplacean principle of insufficient reason.

1 *Introduction*

2 *The Principle of Maximum Entropy*

3 *An Apologia for Judy Benjamin*

4 *Conclusion: Entropy and Insufficient Reason*

1 Introduction

Among the most significant developments in formal epistemology to have taken place over the past several decades is the introduction into the subject of information theoretic concepts and techniques. Of the many figures who contributed to this advance, perhaps the most important is the physicist E. T. Jaynes. Jaynes is best known for his proposal to interpret information entropy in epistemic terms as the amount of 'uncertainty' contained in a probability distribution. Based on this interpretation, he argued for the method of maximizing entropy as the uniquely rational way to infer probabilities from partial or incomplete information.

Despite the crucial role that maximum entropy methods plays in a wide variety of statistical applications, Jaynes's principle of maximum entropy has been subject to severe criticism since its first introduction into the philosophical literature.[1] This criticism is based, in part, on the fact that in certain settings the principle yields counterintuitive results which would appear to conflict with Jaynes's own informal characterization of the maximum entropy distribution as 'minimally informative' or 'maximally non-committal'. The most well-known example of such a counterintuitive result arises in the context of the so-called Judy Benjamin Problem.[2] This problem turns on the apparently puzzling fact that, on the basis of information relating an event's conditional probability, the maximum entropy distribution will almost always assign to the event conditionalized on a probability strictly less than that assigned to it by the uniform distribution.

Some authors have taken the Judy Benjamin problem to provide sufficient grounds for rejecting the principle of maximum entropy and its associated methods.[3] In this paper, I will argue that such a response is unwarranted. The philosophical moral of the Judy Benjamin problem is not that maximum entropy methods have no role to play in a general theory of probabilistic reasoning, but rather that, when employing such methods, care must be taken to ensure that the epistemological prerequisites for their legitimate application have been met.

A detailed analysis of the Judy Benjamin problem can thus shed light on the true justificatory basis of the principle of maximum entropy. Such clarity, in my view, is much needed at present, for the grounds of the principle have been greatly obscured not only by the misguided criticisms of its opponents, but also by the rhetoric of its more outspoken defenders, who conceive of the principle in grandiose terms as supplying the basis of a new and more robust form of 'objective' Bayesianism. Even Jaynes himself, on certain occasions, must be counted among this latter group of misleadingly aggrandizing advocates for the principle. For, Jaynes often sought to portray the principle as a common sense precept of rationality, belying the extent to which its correct application relies upon substantial theoretical knowledge.

The plan of the paper is as follows. In Section 2, I introduce the principle of maximum entropy, emphasizing Jaynes's own characterization of it as a generalization of the Laplacean principle of insufficient reason. In Section 3, I describe the counterintuitive consequences of the principle that give rise to the Judy Benjamin problem. Instead of trying to solve the problem by explaining how these consequences can be avoided, I instead offer a way of making sense of Judy Benjamin's seemingly odd behaviour. This alternative rationalization is based on the assumption that, in applying the principle of maximum entropy, Judy Benjamin is acting out of a concern to exercise a certain brand of epistemic charity toward her informant, one which takes the form of an assumption on the part of Judy Benjamin that her informant's evidential report leaves out no relevant information. Such an account differs substantially from the usual depiction of Judy Benjamin as a conservative agent whose sole aim is to commit herself to as little as possible within the constraints set by her available evidence. In Section 4, I describe how this alternative construal of Judy Benjamin's epistemic motivations can lead to a clearer understanding of the true epistemological grounds of the principle of maximum entropy, and, in particular, shed light on the nature of the relationship between the principle of maximum

[1] Early criticisms of the principle of maximum entropy can be found in (Friedman and Shimony[1971]; Shimony [1973]; Seidenfeld [1979]; Dias and Shimony [1981]).

[2] The Judy Benjamin problem first appeared in (van Fraassen [1981]). For van Fraassen's own analysis of the problem, see (van Fraassen *et al.* [1986]).

[3] Critiques of the principle of maximum entropy that are based on either the Judy Benjamin problem or other closely related problems can be found in (Seidenfeld [1987]; Grove and Halpern[1997]; Gaifman and Vasudevan [2012]). More recent proposals for how to resolve the Judy Benjamin problem appear in (Douven and Romeijn [2011]; Huisman [2014]).

entropy and the principle of insufficient reason.

2 The Principle of Maximum Entropy

The principle of maximum entropy was first introduced by Jaynes in a series of papers published in 1957 ([1957a], [1957b]) on the foundations of statistical mechanics. The overarching aim of these papers was to provide a principled justification for the use of entropic methods in thermodynamics, as most famously exemplified by the derivation of the Maxwell–Boltzmann probability distribution for the velocities of particles in an ideal gas at thermal equilibrium. Contrary to the prevailing orthodoxy at the time, Jaynes argued that such methods do not rely for their justification on the underlying details of any particular physical theory, but instead derive their warrant directly from certain 'common sense' precepts of rationality.[4] In particular, Jaynes argued that these methods could be justified within a broadly Bayesian framework on the basis of a generalized form of the principle of insufficient reason, which he dubbed the principle of maximum entropy. The aim of this section is to provide a brief introduction to this principle, with an emphasis on Jaynes's own view as to its foundational significance for the classical Bayesian theory of probabilistic reasoning.

Bayesianism, in its classical form, has its origins in the mathematical writings of such early theorists of probability as Bernoulli and Laplace. At the center of the theory is the now famous Bayesian rule of inversion—or, what has since come to be known simply as Bayes's rule. This rule describes how the relative probabilities of a family of statistical hypotheses are to be assessed on the basis of some body of observed evidence. More specifically, if $h_1, \ldots, h_n$ are statistical hypotheses and e a body of observed evidence, then Bayes's rule asserts that the probability of the hypothesis h_k indicated by e is given by its conditional probability:

$$Pr(h_k|e) = \frac{Pr(e|h_k)Pr(h_k)}{\sum_{i=1}^{n} Pr(e|h_i)Pr(h_i)}.$$

The power of this simple formula derives from the fact that our intuitive capacity to assess conditional probabilities is, in many ordinary contexts of inquiry, markedly non-invertible. Thus, for example, while we may be in a position to judge the probability with which a fair coin will land heads on exactly three of the next five tosses, we may have no idea how to judge the degree to which such behaviour is indicative of a coin's being fair. Bayes's rule promises us a principled means for overcoming such epistemic asymmetries and, for this reason, it is regarded by many as a logical principle of the highest importance.

In spite of its theoretical promise, however, Bayes's rule is subject to several important constraints, the most fundamental of which derives from its reliance upon an assessment of the probabilities of the hypotheses absent—or 'prior' to—the collection of the evidence. How exactly are these prior probabilities, $Pr(h_1), \ldots, Pr(h_n)$, to be assessed? The only principle for inferring prior probabilities that was known to Laplace and the classical Bayesians was the so-called principle of insufficient reason, which asserts that when there is no reason to judge any one hypothesis more likely than any other, one ought to assign to each an equal probability.[5]

[4] Many physicists, including Boltzmann himself, took the validity of the derivation of the Maxwell–Boltzmann distribution to rely on a tacit ergodic hypothesis, equating the time and ensemble averages of certain macroscopically measurable quantities in the thermodynamic systems under study (see Goldstein [2001]). Jaynes, on the other hand, did not consider such claims about ergodicity to be at all relevant to the foundations of thermodynamics. On this point, he took himself to be in agreement with Gibbs (see Jaynes [1967]).

[5] This principle of insufficient reason has, of course, been subject to numerous objections over the years. Most of these objections are based on an interpretation of the principle according to which its legitimate application

Now, however one may choose to interpret this principle, it would seem to possess a rather limited scope of applicability. For, there appear to be very few contexts of inquiry, apart from contrived examples involving simple games of chance, in which our prior knowledge of the situation renders all the hypotheses under consideration equally possible. Such limitations of the principle of insufficient reason were obvious even to the classical Bayesians, and, according to Jaynes, it was precisely their failure to provide any more general solution to the problem of prior selection, that rendered the classical Bayesian view susceptible to critique.[6] Thus, Jaynes ([1978], p. 217) writes:

> The only useful results Laplace got came from [. . .] the principle of insufficient reason. That is, of course, not because Laplace failed to understand the generalization [of Bayes's rule to non-uniform prior probabilities] as some have charged [. . .] Rather, Laplace did not have any principle for finding prior probabilities in cases where the prior information fails to render the possibilities 'equally likely' [. . .] The next order of business should have been seeking new and more general principles for determining prior probabilities [. . .] Instead, only fifteen years after Laplace's death, there started a series of increasingly violent attacks on his work. Totally ignoring the successful results they had yielded, Laplace's methods based on [Bayes's rule] were rejected and ridiculed, along with the whole conception of probability expounded by Bernoulli and Laplace.

In Jaynes's view, a rebirth of Bayesianism would thus have to await the discovery of 'new and more general' methods for inferring prior probabilities. Among the first to make significant progress on this score was the statistician Harold Jeffreys. Jeffreys proposed a general rule which was meant to supplement the discrete form of the principle of insufficient reason by providing a systematic procedure for constructing a prior reflecting a state of total ignorance with respect to a continuous space of hypotheses.[7] While Jaynes ([1978], p. 219) criticized Jeffreys for not showing how his rule could be derived as the 'necessary consequence of any compelling desiderata', he endorsed what he took to be the general conceit underlying Jeffreys's reasoning, namely, that our prior probabilities should be 'minimally informative' in the sense that they should have a minimal impact on the updated form of the probability distribution once Bayes's rule has been applied.[8]

requires an agent to be in a state of 'total ignorance' as to which of the hypotheses is true. As the objection goes, since total ignorance is a robust enough state as to be preserved through non-uniform regroupings of the hypotheses, the principle, in its unqualified form, issues in inconsistent recommendations. The *locus classicus* of such 'paradoxes of indifference' is (Keynes [1921], Chapter 4). For Jaynes's proposed solution to a specific paradox of this sort, see (Jaynes [1973]).

[6] With regard to the scope of applicability of the principle of insufficient reason, Bernoulli ([2006], pp. 326–7) wrote:

> [Such *a priori* methods] only extremely rarely succeed, and hardly ever anywhere except games of chance which their first inventors, desiring to make them fair, took pains to establish in such a way that the [. . .] cases themselves occurred with the same facility [. . .] But what mortal, I ask, may determine, for example, the number of diseases as if they were just as many cases, which may invade at any age the innumerable parts of the human body and which imply our death? And who can determine how much more easily one disease may kill than another—the plague compared to dropsy, dropsy compared to fever? In these and similar situations [. . .] it would clearly be mad to want to learn anything in this way.

[7] See (Jeffreys [1961], Chapter 3). A comprehensive survey of methods for selecting non-informative priors containing a detailed discussion of Jeffreys's rule is given by Kass and Wasserman ([1996]).

[8] Jeffreys himself argued for his rule on the grounds that the resulting priors were invariant under

To make this idea of a minimally informative prior precise, Jaynes would borrow from the work of Claude Shannon, a mathematician and engineer at Bell Labs, whose 1948 monograph titled *A Mathematical Theory of Communication* is one of the founding documents of information theory.[9] In this pioneering work, Shannon set out to provide a mathematical model of what he termed a 'communication system', that is, a system by which a randomly produced message is transmitted to a receiving terminal over some (possibly noisy) channel. At the heart of Shannon's theory was the simple but profound observation that the particular probability distribution characterizing the chance mechanism by which a communication system outputs its messages has a direct impact on how efficiently those messages can be encoded for transmission across the channel. If, for example, this distribution is more uniform, only slight gains in efficiency can be obtained through the use of shorter codes for the more frequently outputted symbols. If, on the other hand, the distribution is more sharply peaked, so that the probability of outputting certain symbols is much higher than that of outputting others, then the use of shorter codes for the more frequently occurring symbols can result in much more dramatic gains in efficiency.[10]

Among the most important contributions of Shannon's work was a precise specification of the theoretical limit to how efficient a lossless code might be, given the probabilities with which various symbols are outputted from an information source. In this way, Shannon was able to associate with every probability function a certain number, referred to as its information entropy, proportional to the number of bits per symbol required on average to encode a message produced by a communication system characterized by this probability function. Specifically, if Pr is a probability function on an algebra whose atoms are $h_1, \ldots, h_n$, then the information entropy of Pr is given by:

$$S(Pr) = -\sum_{i=1}^{n} Pr(h_i) \log Pr(h_i)$$

In Shannon's measure of information entropy, Jaynes saw the potential basis for a fully general criterion for Bayesian prior selection. The crucial step in Jaynes's analysis was a reconceptualization of the probabilities appearing in Shannon's theory, in subjective or epistemic terms. In particular, Jaynes proposed to interpret these probabilities not as the long-run relative frequencies with which messages are produced by the source, but instead as a description of the state of partial knowledge of an agent who is concerned to predict what this outputted message will be. Jaynes ([1957a], p. 622) argued that once the probabilities are interpreted this way, Shannon's formula for information entropy can be reconstrued as the amount of 'uncertainty' that characterizes an agent whose epistemic state is represented by a given probability

reparametrization. It can, however, be shown that when Jeffreys's rule is applied in ordinary cases the resulting priors maximize an asymptotic form of relative entropy (Clarke-Barron [1994]). In this sense, then, Jeffreys's rule represents a particular instance of the more general method of maximizing entropy.

[9] See (Shannon [1948]). A historical account of the early developments in information theory is given by Pierce ([1980], Chapters 2–3).

[10] Suppose, for example, that the messages produced by the source consist of strings of symbols from the alphabet $\{A, B, C\}$ and that we wish to communicate these messages to a digital computer. This requires that each symbol outputted from the source be encoded into a binary string, the elements of which (bits) may be represented by either a 0 or a 1. Now, if it is discovered that the letter B is much more likely to occur in messages from the source than either of the letters A or C, we can dramatically improve the efficiency of our code—that is, we can reduce, on average, the length of our messages to the computer—by adopting an encoding scheme which assigns to the letter B a shorter binary sequence. Indeed, if we think of a natural language like English as an autonomic code (that is, a code in which the code language and the source language are the same), our use of shorter words, like 'and', 'the' and 'but', to encode those words which appear most often in our speech can be given a similar justification in terms of the improved communicative efficiency afforded by such a practice.

distribution[11]:

> The great advance provided by information theory lies in the discovery that there is a unique, unambiguous criterion for the 'amount of uncertainty' represented by a discrete probability distribution, which agrees with our intuitive notions that a broad distribution represents more uncertainty than does a sharply peaked one, and satisfies all other conditions which make it reasonable [. . .] Since this is just the expression for entropy as found in statistical mechanics [. . .] henceforth we will consider the terms 'entropy' and 'uncertainty' as synonymous.

If, as Jaynes suggests, entropy is simply a measure of uncertainty, then an agent whose epistemic state can be represented by the probability distribution that maximizes entropy under the constraints imposed by that agent's evidence can be described as maximally uncertain as to which of the hypotheses under consideration is true. Such an attitude of maximal uncertainty, according to Jaynes, reflects 'the only unbiased assignment [of probabilities] we can make',[12] for any other probability distribution would contain more information than what is needed to account for the evidence. For this reason, Jaynes was led to advance the following principle of probabilistic reasoning:

> Principle of Maximum Entropy: If the set of probabilities consistent with one's evidence is Δ, then, on the basis of such evidence, one ought to adopt a probability distribution $Pr \in \Delta$ satisfying:
>
> $$H(Pr) = \max\{H(P) : P \in \Delta\}.$$

As Jaynes pointed out, the principle of maximum entropy reduces to the principle of insufficient reason in the specific case in which the agent's evidence imposes no constraints on his choice of probabilities. This is because, among all probability distributions, the uniform distribution has maximal entropy. Over and above this limiting case, however, the principle determines a unique prior probability distribution in many instances in which the agent's prior knowledge imposes non-trivial constraints on this distribution's form.[13] Suppose, for example, that an agent knows that the mean value of a certain random variable X is x, that is,

$$\sum_{i=1}^{n} Pr(h_i)X(h_i) = x \tag{1}$$

[11] See (Jaynes [1983b], pp. 233–7). The conceptual connection between information entropy and uncertainty can be understood along roughly the following lines: if an agent were reasonably certain as to what the outcome of a given chance process will be, then, on the basis of his knowledge, he ought to be able to devise a fairly efficient language in which to reliably report the outcomes of this process to another—or, equivalently, he ought to be able to insure against the loss of information at a relatively cheap rate. If, on the other hand, the agent were more uncertain as to what the outcome of the process will be, he should expect this language to be less efficient—or the insurance to be more costly.

[12] (Jaynes [1957a], p. 623).

[13] Since the entropy $H(P)$ is a strictly concave function on the vector space of probabilities, the principle of maximum entropy will recommend a unique probability distribution provided the set of evidentially admissible probabilities Δ is convex. In particular, if the agent's knowledge imposes linear constraints on the probability (that is, constraints which can be expressed as one or more linear equations in $P(h_1), \ldots, P(h_n)$), the principle will issue in a unique prior. In what follows, the only constraints that we will consider will be linear constraints, so that we may speak unequivocally of the maximum entropy distribution.

On the basis of this evidence, the principle of maximum entropy will instruct the agent to adopt a prior probability distribution with an exponential form:

$$Pr(h_i) \propto e^{-\lambda X(h_i)},$$

where λ is a constant determined by Equation (1).

Jaynes viewed the principle of maximum entropy as supplying the missing criterion for prior selection needed in order to revitalize the Bayesian programme. By utilizing this principle, Bayesianism could be transformed from a mere tool for analysing simple games of chance into a powerful and more comprehensive methodology capable of underwriting even sophisticated modes of reasoning such as those employed in the statistical sciences. Notwithstanding such lofty ambitions, Jaynes himself conceived of the principle of maximum entropy as deriving its rational warrant from very modest epistemological assumptions. According to Jaynes, maximum entropy reasoning is nothing more than a quantitatively precise implementation of the 'common sense' maxim to adopt beliefs which leave one as uncertain as one's evidence will permit. Thus, the principle's normative force, even when applied in scientific contexts, does not derive from any special theoretical knowledge, but instead from the conservative ideal of an epistemic agent who is willing to believe only what is needed to account for his evidence. It is this attempt by Jaynes to provide an *a priori* justification of the principle of maximum entropy by appeal to the policy of epistemic conservatism that has invited criticism of the principle in the more recent philosophical literature.[14] For in certain simple contexts of inquiry the principle issues in recommendations which are anything but conservative.

3 An Apologia for Judy Benjamin

The most well-known objection to the principle of maximum entropy is based on the Judy Benjamin problem, first introduced by van Fraassen ([1981]). The problem arises in the context of an imagined scenario inspired by a scene from the film *Private Benjamin* starring Goldie Hawn. In this scenario, Private Judy Benjamin and her platoon are participating in war games. The region in which the war games are conducted is divided into two zones: the blue army zone, which is friendly territory, and the red army zone, which is enemy territory. Each of these zones is further divided into two sub-regions, one containing the controlling army's headquarters and the other containing its second company's camp.

Judy Benjamin and her platoon are air-dropped into an area which they are asked to reconnoiter. She is given a map which she cannot decipher and after a short while she and her entire platoon are hopelessly lost. She radios to headquarters seeking additional information as to her whereabouts, and her contact at headquarters responds with the following message: 'If you are in the red army zone, then the probability is ¾ that you are in the enemy's headquarters region'. How should Judy Benjamin respond to this message?

There are four possibilities to consider:

R_1 : Judy Benjamin is somewhere in the red army headquarters region.

[14] Jaynes's predilection to view the principle of maximum entropy as an *a priori* principle of rationality can be clearly gleaned from the significance he attaches to the representation theorem of Shannon, which shows that (up to a multiplicative constant) the entropy is the unique measure satisfying certain mathematical properties. Jaynes ([1957a], p. 630) characterized Shannon's result as a derivation of the entropy measure from 'elementary conditions of consistency'. For other attempts to provide an axiomatic justification of the principle of maximum entropy, see (Cox [2001], Chapter 2; Shore and Johnson [1980]; Csiszar [1991]; Paris and Vencovská [1990]).

R_2 : Judy Benjamin is somewhere in the red army second company region.

B_1 : Judy Benjamin is somewhere in the blue army headquarters region.

B_2 : Judy Benajmin is somewhere in the blue army second company region.

Let B be the event that Judy's platoon is located somewhere in the blue army zone and R be the event that it is located somewhere in the red army zone (that is, $B = B_1 \vee B_2$ and $R = R_1 \vee R_2$). When Judy contacts headquarters, she is told that if she is in the red army zone, the probability that she is in the red army headquarters region is 3/4. In other words, she is told that

$$P(R_1|R) = 3/4. \tag{2}$$

Among all the probability distributions satisfying this constraint, which one should Judy Benjamin adopt? Even without appealing to any formal principles of probabilistic reasoning, one may already surmise what a conservative answer to this question should look like. Since, on the one hand, the evidence provides Judy Benjamin with no relevant information as to whether she is in the blue or the red army zone, it seems plausible to suppose that she ought to conclude that the former possibility is just as likely to obtain as the latter. Likewise, on the assumption that Judy Benjamin is somewhere in the blue army zone, she ought to conclude that she is as likely to be in the army's headquarters region as she is to be in its second company's region. All told then, based only on the evidence at her disposal, Judy Benjamin's probabilities ought to satisfy the following two constraints:

$$P(R) = P(B), \tag{3}$$

$$P(B_1|B) = P(B_2|B). \tag{4}$$

Conditions (2)–(4) determine a unique probability distribution, which we shall henceforth refer to as Pr_{CON}. The mathematical fact that underwrites the Judy Benjamin problem is that Pr_{CON} turns out not to be the probability distribution that maximizes entropy under the constraint described by condition (1).[15] Let us refer to this latter, entropy maximizing probability distribution as Pr_{ME}. While Pr_{ME} does satisfy Condition (4), it does not satisfy Condition (3). More specifically, the maximum entropy distribution assigns to the event R a probability strictly less than that which it assigns to the event B (see Figure 1).

This is a rather puzzling fact, for why should information that only imposes constraints on Judy Benjamin's conditional probabilities lead her to conclude unconditionally that she is less

[15] van Fraassen ([1981]) presents the Judy Benjamin problem not as a difficulty for the principle of maximum entropy, viewed as a method for prior selection, but rather as a challenge to the principle of cross-entropy maximization, viewed as a method for updating a prior. In one sense, the difference is not important. For, the maximum entropy distribution is just the cross-entropy maximum update of the uniform distribution. Consequently, the Judy Benjamin problem can seen as a criticism of the principle of maximum entropy provided we interpret Judy Benjamin's prior probabilities as the result of 'updating' a uniform prior. Still, there is an important conceptual difference between choosing a prior and updating a prior. This is reflected in the fact that the analysis of the Judy Benjamin problem offered in this paper does not generalize in any straightforward way to the case of updating a non-uniform prior. In this regard, it is perhaps worth noting that Jaynes himself never proposed the principle of maximum entropy be used for the purpose of updating prior probabilities. On the only occasion on which Jaynes makes explicit reference to relative entropy, he introduces it as an alternative to the χ^2 statistic, as a measure of the 'goodness of fit' between a statistical hypothesis and the empirical measure obtained from a data sample. This appeal to the relative entropy is clearly very different from its use in updating probabilities, where it is taken as a measure of the distance between statistical hypotheses themselves.

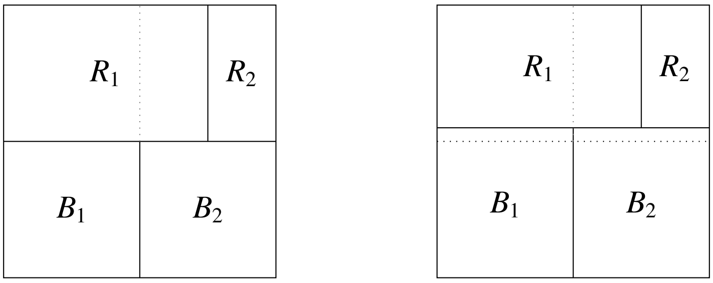


Figure 1: The area-chart on the left depicts the probability function Pr_{CON}. The area-chart on the right depicts the probability function, Pr_{ME}, that maximizes entropy under the constraint $P(R_1|R) = 3/4$. Note that $Pr_{\text{ME}}(R) < Pr_{\text{ME}}(B)$.

likely than not to be in the red army zone? In response to such seemingly bizarre reasoning, van Fraassen ([1981], p. 379) remarks:

> It is hard not to speculate that the dangerous implications of being in the enemy's headquarters area, are causing Judy Benjamin to indulge in wishful thinking, her indulgence becoming stronger as her conditional estimate of the danger increases.

Van Fraassen's diagnosis of Judy Benjamin is surely meant to be tongue-in-cheek. Nevertheless, it is an instructive exercise to consider why the description of Judy Benjamin as engaged in 'wishful thinking' does not adequately capture what is objectionable about her appeal to the principle of maximum entropy. For one thing, such a description of Judy Benjamin relies on various assumptions about her practical aims which are inessential to a recognition of the apparent doxastic defect in her reasoning.[16] Even if, for some reason, Judy Benjamin desperately hoped to find herself in enemy territory (perhaps with the aim of capturing the enemy's headquarters), she would still be subject to criticism insofar as her probabilities appear to commit her to more than what the evidence requires. Since this critique is an epistemological and not a practical one, it is irrelevant whether such epistemic excesses lead Judy Benjamin to adopt an unjustifiably pessimistic or optimistic attitude toward her situation.

[16] While Judy Benjamin's practical aims are irrelevant to the problem, we can further highlight the counterintuitive nature of her behaviour by embedding the problem within a decision-theoretic context. Suppose that Judy Benjamin deciphers her map enough to know how to reach the border dividing the red army from the blue army zone. She knows that moving her platoon towards the border will either move them from one army's headquarters region to the other's, or from one army's second company's region to the other's. If we suppose that Judy Benjamin would like to find her way back to friendly territory where she and her platoon will be safe, and that she assesses the losses of being discovered in the enemy's headquarters region as greater than those of her being discovered in the enemy's second company's region, then the decision problem confronting Judy can be represented by the following table (where $0 < x < 1$):

	R_1	R_2	B_1	B_2
Stay	-1	$-x$	0	0
Move	0	0	-1	$-x$

It is not hard to see that there exists an $x < 1$ such that, with respect to the maximum entropy probability distribution, the expected utility of Stay is greater than the expected utility of Move. But this means that Judy Benjamin may treat the report from headquarters, informing her that if she is in enemy territory, she is more likely to be in the more dangerous of the two sub-regions, as providing her with a reason to stay put.

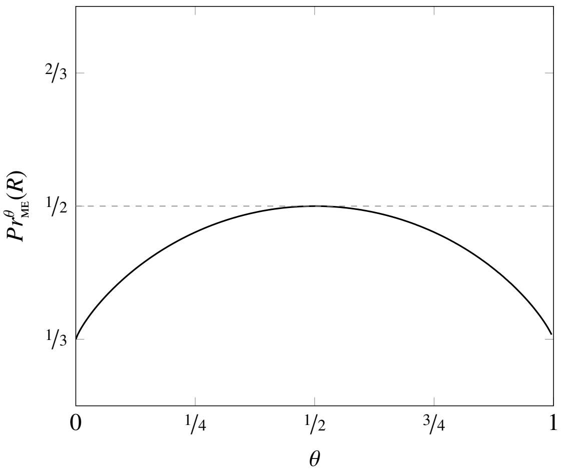


Figure 2: $Pr^{\theta}_{\text{ME}}(R)$ as a function of θ. $Pr^{\theta}_{\text{ME}}(R)$ obtains its maximum value of $1/2$ when $\theta = 1/2$, and is a strictly decreasing function in $|\theta - 1/2|$.

Secondly, the claim that Judy Benjamin's indulgence in wishful thinking grows 'stronger as her conditional estimate of the danger increases' carries with it the suggestion that the greater the conditional probability she assigns to R_1 given R, the less likely she will judge it to be that she is somewhere in the red army zone. This, however, is not true. For any $\theta \in [0, 1]$, let Pr^{θ}_{ME} be the probability function that maximizes entropy under the constraint, $P(R_1|R) = \theta$.[17] Then, $Pr^{\theta}_{\text{ME}}(R)$ is not a decreasing function in θ. Instead, it obtains its maximum value of $1/2$ when $\theta = 1/2$ and decreases as θ moves farther away from this value in either direction, obtaining its minimum value of $1/3$ at the extremes, $\theta = 0$ and $\theta = 1$ (see Figure 2).[18] Thus, regardless of the value of $P(R_1|R)$ reported to Judy Benjamin by her contact at headquarters, choosing her probabilities in accordance with the principle of maximum entropy will always lead her to conclude that she is at least as likely to be somewhere in the blue army zone as she is to be in the red army zone. Moreover, provided the reported value is not equal to $1/2$, this inequality will be strict.[19]

If Judy Benjamin is not engaged in wishful thinking, how are we to understand the logic behind her reasoning? Recall that the principle of maximum entropy can be interpreted as the recommendation to make one's probabilities as uniform as possible within the constraints set by the available evidence. Now, when Judy Benjamin is informed of the conditional probability of her being in the red army's headquarters region given that she is somewhere in the red army zone, her new information requires her to introduce certain non-uniformities into her

[17] We stipulate that no probability function satisfies such a constraint that assigns to R a probability of 0.

[18] More specifically:

$$Pr^{\theta}_{\text{ME}}(R) = \begin{cases} \frac{\beta(\theta)}{1+\beta(\theta)} & \text{if } \theta \in (0,1) \\ 1/3 & \text{if } \theta = 0 \text{ or } \theta = 1 \end{cases},$$

where

$$\beta(\theta) = \frac{(1-\theta)^{\theta-1}}{2\theta^{\theta}}, \quad \text{for } \theta \in (0,1).$$

[19] In general, if E and F are events in a finite Boolean algebra, then the probability function Pr on that algebra which maximizes entropy under the constraint $P(E|F) = \theta$ (where $\theta \in [0, 1]$) will satisfy the inequality $Pr(F) \leq Pr_0(F)$, where Pr_0 is the uniform probability function. Moreover this inequality will be strict unless $\theta = Pr_0(F)$. The first proof of this claim is given by (Seidenfeld [1987], p. 282, app. B, res. 4).

probability distribution. Note, however, that these non-uniformities apply only to the relative probabilities of those possible outcomes that are consistent with the conditioning event, in this case, the event R. It therefore stands to reason that Judy Benjamin could minimize the overall effect of such non-uniformities by confining them to an event of lower probability, or, equivalently, by assigning a lower probability to R.[20]

In applying the principle of maximum entropy, Judy Benjamin can thus be viewed as attempting to marginalize the informational impact of the report she receives from headquarters by assigning a lower probability to the event that this report will be impactful. Whatever else may be said in favour of such reasoning, it is certainly not 'conservative', for, in this case, the report has its maximal impact just in case Judy Benjamin is somewhere in the red army zone; but the report itself does not contain any information relevant for assessing how likely this is to be the case.[21] In more general terms, the difficulty that lies at the heart of the Judy Benjamin problem derives from the fact that not all probabilistic information is intuitively relevant for assessing the probability that this information will itself turn out have a more or less significant impact on one's probabilities. Information concerning conditional probabilities is a clear case in point. Intuitively, such information is irrelevant for assessing how likely it is that the conditioning event will occur, and yet the occurrence of this event is precisely the condition under which this information has its greatest impact.[22]

The standard response to the Judy Benjamin problem has been to reject the principle of maximum entropy in favour of some alternative methodology that better reflects our intuitions as to what a conservative response to the evidence should look like.[23] But is this reaction warranted? Does the Judy Benjamin problem provide us with grounds for rejecting the principle of maximum entropy, or can the principle be somehow salvaged? Note that the Judy Benjamin problem can only serve as a direct critique of the principle of maximum entropy provided the latter principle is understood as deriving its rational warrant from the conservative maxim to remain as non-committal as possible within the constraints set by the available evidence. But is this the right way to view the principle? On the one hand, Jaynes himself certainly seemed to have conceived of the principle in this way. This much is clear from his identification of entropy with uncertainty and his frequent characterization of the maximum entropy distribution

[20] This is not a proof. There are further details that must be worked out since there is a 'cost' in information associated with Judy Benjamin's non-uniform assignment of probabilities to the events R and B. For a detailed account of the cost-balancing arguments in the case of constraints on an agent's conditional probabilities, see (Gaifman and Vasudevan [2012], Section 5).

[21] Williamson ([2010], p. 77, Footnote 4) concedes that the maximum entropy distribution is not maximally conservative. He argues that it is nonetheless maximally 'equivocal', where equivocality is measured by proximity to an 'equivocator' function—in this case, the uniform distribution. What Williamson does not explain is why should we value 'equivocation' (which he treats as a primitive rational norm) once it has been acknowledged that the most conservative attitude to adopt towards our evidence is not one of maximal equivocation.

[22] Information about conditional probabilities provides only one example of this sort of information. Consider, for example, a report that describes the expected (average) value of a random variable X which takes values in the set $\{1, 2, 3, 4\}$. This report will have a more significant impact on one's probabilities if it were to turn out that the value of X is in the set $\{2, 3\}$, since more significant changes in the expected value of X can be effected through smaller changes in the probabilities assigned to the extremal values 1 and 4 than to the central values 2 and 3. As a result of this 'torque'-like effect, maximizing entropy under expected-value constraints will lead to a lowering of the probability of more central values of X at the expense of extremal values despite the fact that this seems not to be justified by the evidence (see Shimony [1973]).

[23] Such an alternative methodology should, in particular, lead Judy Benjamin to adopt the distribution $Pr_{\textsc{con}}$ in response to the report she receives from headquarters. For a few examples of such conservative methodologies, see (van Fraassen *et al.* [1986]). An interesting recent proposal for a conservative methodology, described by Douven and Romeijn ([2011]), is the proposal to maximize 'inverse relative entropy', which is obtained from the usual relative entropy by swapping the prior and posterior probabilities.

as maximally non-committal. At the same time, by training, Jaynes was a physicist and not a philosopher, and, as a result, his focus tended to be directed more toward the applications of the method than its epistemological foundations. Therefore, it is perhaps wise not to simply accept on faith Jaynes's own account of the justification of the principle of maximum entropy but to instead approach his more overtly epistemological pronouncements with some initial degree of scepticism.

In the remainder of this section, we will explore the possibility that, notwithstanding Jaynes's claims to the contrary, the principle of maximum entropy is not an epistemically conservative principle, but that it instead derives its warrant from considerations of an altogether different sort. Accordingly, we will no longer take it for granted that Judy Benjamin's sole concern in processing the report from headquarters is to remain maximally uncertain in the face of new evidence. Rather, we will start from the assumption that Judy Benjamin's apparently puzzling reaction to the report is, in fact, reasonable given her epistemic aims and motives. Taking this as our starting point, we will then proceed to inquire as to what these aims and motives might be. We will focus, in particular, on the question of how to make sense of the fact that, on the basis of the report from headquarters, Judy Benjamin concludes that she is less likely than not to be in the red army zone.

How might we rationalize this conclusion on the part of Judy Benjamin? The only assumption that we have made so far about how Judy Benjamin conceives of the report from headquarters is that she regards the information it contains as reliable in the sense of imposing a constraint on her choice of probabilities. But we have not yet attributed to her any view as to why her informant decided to provide her with this specific information. Why, that is, did he choose to inform her of the conditional probability of R_1 given R? Of course, no specific views concerning the intentions of her informant are imputed to Judy Benjamin in the original description of the scenario. Indeed, it may be that her view is that the report simply contains whatever information happened to be available at the time. Nevertheless, let us suppose, for the sake of the argument, that this is not the case; that Judy Benjamin does not believe that the information contained in the report was chosen arbitrarily, but instead that this information was communicated to her precisely because it contains all the probabilistically relevant information relating to her situation.

Can such an assumption help to make sense of Judy Benjamin's seemingly puzzling behaviour? In order to answer this question, we must first state more clearly what it means for Judy Benjamin to assume that the report contains 'all the probabilistically relevant information' relating to her situation. To begin with, let us suppose that Judy Benjamin's informant interprets her request for additional information in broad, theoretical terms, not just as a request for specific information relating to her current whereabouts, but, more generally, as a request for information relating to the underlying chance process by which she and her platoon ended up in the region they now occupy. More specifically, let us suppose that both Judy Benjamin and her informant view her platoon's present location as the outcome of a single trial in a repeatable chance process consisting of several airdrops into the region where the war games are taking place. Let X_n ($n = 1, 2, \ldots, N$) be the outcome of the nth trial of this process. Then, X_n is a random variable whose value is in the set $\{R_1, R_2, B_1, B_2\}$. When Judy Benjamin contacts headquarters requesting additional information, we will assume that her informant takes her to be requesting information relating to the joint probability distribution for the variables $X_1, \ldots, X_N$.

In response to Judy Benjamin's request for additional information, her informant replies that if Judy Benjamin is somewhere in the red army zone, then the probability that she in the red army's headquarters region is 3/4. We will suppose that Judy Benjamin interprets this report as providing a partial description of the outcome of the chance process described by the variables

$X_1, \ldots, X_N$. In particular, if Θ is the proportion of landings in the red army's headquarters region to landings in the red army zone, then the report from headquarters contains the information that $\Theta = 3/4$, that is, that exactly $3/4$ of the variables $X_1, \ldots, X_N$, which take values in the set $\{R_1, R_2\}$, take the value R_1.

Upon receiving this report, Judy Benjamin may well ask herself why it is that her informant at headquarters chose to relate to her of the value of Θ. Was it because this was the only information at his disposal, or was her informant's decision to report this value a deliberate one? It is here that we will ascribe to Judy Benjamin the further belief that she was informed of the value of Θ precisely because this quantity contains all the probabilistically relevant information related to her situation. This belief imposes additional constraints on the joint probability distribution for $X_1, \ldots, X_n$ since it implies that as long as two possible outcomes agree in the value of Θ, any further differences between them may be ignored for the purpose of assessing their relative probabilities. Equivalently, any two distinct outcomes that are consistent with the report agree in all probabilistically relevant respects and so should be assigned the same probability.[24]

Let us restate this constraint in more precise terms. Let Ω be the set of all possible outcomes of the chance process described by $X_1, \ldots, X_N$. Then, Ω can be represented by the set of all N-length sequences of members of the set $\{R_1, R_2, B_1, B_2\}$. For any such sequence, $x_1, \ldots, x_N$, we write $\Theta(x_1, \ldots, x_N)$ for the ratio of R_1's to R_1's or R_2's among the x_n's. Judy Benjamin's assumption that Θ contains all the probabilistically relevant information related to her situation, can then be captured by the following constraint. For all $\omega, \omega' \in \Omega$:

$$\text{If } \Theta(\omega) = \Theta(\omega'), \text{ then } Pr(\omega) = Pr(\omega'). \tag{5}$$

This assumption turns out to be all that is needed in order to explain Judy Benjamin's puzzling behaviour. For, as we shall see, given this assumption, it follows from the reported fact that $\Theta = 3/4$, that the probability, on any given airdrop, of the platoon landing somewhere in the red army zone is strictly less than $1/2$. In other words, it follows directly from Condition (5) that, for any given n:

$$Pr\,(X_n = R \,|\, \Theta = 3/4) < 1/2.^{25}$$

To see why this is so will require some explanation.[26] We first note that, since rearranging the variables $X_1, \ldots, X_N$ has no effect on the value of Θ, Condition (5) implies that these variables are 'exchangeable' in the sense that their joint distribution is invariant under arbitrary permutations. This, in turn, implies that their joint distribution can be represented by the following urn model.[27] Consider the collection of all possible urns containing N balls, each of which is either marked R_1, R_2, B_1, or B_2. From this collection, choose an urn U at random with probability $p(r_1, r_2, b_1, b_2)$, where r_1 is the number of balls marked R_1 in U; r_2 the number of balls marked R_2 in U; and so forth. Draw the balls from the selected urn at random one at a time without replacement, and let the value of the variable X_n be determined by the mark on the nth ball drawn from the urn.

[24] It is worth emphasizing that, in what follows, N is treated as a fixed parameter known to both Judy Benjamin and her informant. Thus, when we say that her informant's report contains all the relevant information related to her situation, we mean by this that the report contains all the information relevant to assessing the relative probabilities of any two possible outcomes of an experiment consisting of a fixed number N of repeated trials.

[26] A reader who is uninterested in the mathematical details may skip ahead to the paragraph beginning 'Thus, we have shown...'.

[27] For a concise and elegant proof of de Finetti's theorem which relies on the connection between exchangability and urn models of the sort described above, see (Heath and Sudderth [1976]).

Now, consider an urn U characterized by the numbers r_1, r_2, b_1, b_2.[28] The number of distinct possible ways in which all N balls can be drawn from this urn is given by the multinomial coefficient:

$$\binom{N}{r_1\,,\ r_2\,,\ b_1\,,\ b_2} = \frac{N!}{r_1!r_2!b_1!b_2!}.$$

Since each of these possible outcomes is equally probable, if we let ω be any one such outcome, it follows from the above urn model that

$$Pr(\omega) = \left(\frac{1}{\binom{N}{r_1\,,\,r_2\,,\,b_1\,,\,b_2}}\right) p(r_1, r_2, b_1, b_2).$$

Putting $r = r_1 + r_2$ and $\theta = \frac{r_1}{r}$, this can be rewritten as:

$$Pr(\omega) = \left(\frac{1}{\binom{N}{r\theta\,,\,r(1-\theta)\,,\,b_1\,,\,N-(r+b_1)}}\right) q(r, \theta, b_1), \tag{6}$$

where

$$q(r, \theta, b_1) = p(r\theta, r(1-\theta), b_1, N - (r + b_1)).$$

Now, by Condition (5), for fixed θ, $Pr(\omega)$ has a constant value. It thus follows from Equation (6) that

$$q(r, \theta, b_1) \propto \binom{N}{r\theta\,,\ r(1-\theta)\,,\ b_1\,,\ N-(r+b_1)},$$

where the constant of proportionality is a constant determined by θ.

Now, consider all the urns in which $\Theta = \theta$. Let d be the denominator of θ, when θ is expressed in lowest terms. If N_R is the number of balls marked either R_1 or R_2 in a given urn, then among those urns in which $\Theta = \theta$, N_R can take any value in the set:

$$\{d, 2d, \ldots, Kd\},$$

where K is the largest integer such that $Kd \leq N$. The probability of selecting from among these urns, one in which $N_R = kd$ is therefore:

$$\begin{aligned}
Pr(N_R = kd|\Theta = \theta) &= \sum_{b_1=0}^{N-kd} q(kd, \theta, b_1) \\
&\propto \sum_{b_1=0}^{N-kd} \binom{N}{kd\theta\,,\ kd(1-\theta)\,,\ b_1\,,\ N-(kd+b_1)} \\
&= \sum_{b_1=0}^{N-kd} \binom{N}{kd\theta\,,\ kd(1-\theta)\,,\ N-kd} \binom{N-kd}{b_1} \\
&= \binom{N}{kd\theta\,,\ kd(1-\theta)\,,\ N-kd} \sum_{b_1=0}^{N-kd} \binom{N-kd}{b_1} \\
&= \binom{N}{kd\theta\,,\ kd(1-\theta)\,,\ N-kd} 2^{N-kd}.
\end{aligned}$$

[28] Such basic combinatorial facts relating to urn models and other sampling models can be found in any elementary textbook on probability, for example, (Feller [1968], Chapter 2).

Normalizing, we have:

$$Pr(N_R = kd|\Theta = \theta) \quad = \quad \frac{\binom{N}{kd\theta\,,\,kd(1-\theta)\,,\,N-kd}2^{-kd}}{\sum_{j=1}^{K}\binom{N}{jd\theta\,,\,jd(1-\theta)\,,\,N-jd}2^{-jd}}. \tag{7}$$

Now, once an urn has been selected, the probability that, on any given drawing, a ball marked either R_1 or R_2 will be drawn is just $\frac{N_R}{N}$. Thus, all told, it follows from Condition (5) that

$$\begin{aligned} Pr(X_n = R|\Theta = \theta) \quad &= \quad \sum_{k=1}^{K} Pr(X_n = R|N_R = kd, \Theta = \theta) \times Pr(N_R = kd|\Theta = \theta) \\ &= \quad \sum_{k=1}^{K}\left(\frac{kd}{N}\right) Pr(N_R = kd|\Theta = \theta) \\ &= \quad \frac{\sum_{k=1}^{K}\left(\frac{kd}{N}\right)\binom{N}{kd\theta\,,\,kd(1-\theta)\,,\,N-kd}2^{-kd}}{\sum_{k=1}^{K}\binom{N}{kd\theta\,,\,kd(1-\theta)\,,\,N-kd}2^{-kd}}. \end{aligned} \tag{8}$$

The crucial point to note about this equation is that, for large enough N and for any $\theta \neq 1/2$, the quantity on the right-hand side of the equation is strictly less than $1/2$. More specifically, this inequality holds whenever $N > 2d$.[29] Thus, provided the value of Θ reported to Judy Benjamin by her informant is not equal to $1/2$, the belief that this information contains all the relevant probabilistic information related to her situation is sufficient to infer that she is less likely than not to be in the red army zone.[30]

For the sake of illustration, let us confirm this for the specific case in which $\theta = 3/4$ (so that $d = 4$) and $N = 4K$, for some $K \geq 3$. In this case, we have to establish the following inequality:

$$\frac{\sum_{k=1}^{K}\left(\frac{k}{K}\right)\binom{4K}{3k\,,\,k\,,\,4(K-k)}2^{-4k}}{\sum_{k=1}^{K}\binom{4K}{3k\,,\,k\,,\,4(K-k)}2^{-4k}} < \frac{1}{2}.$$

This can be rewritten:

$$\sum_{k=1}^{K}\left(1 - \frac{2k}{K}\right)\binom{4K}{3k\,,\,k\,,\,4(K-k)}2^{-4k} \quad > \quad 0.$$

By rearranging the terms in the left-hand sum, we can express this inequality as follows:

$$\sum_{1\leq k<K/2}\left(1 - \frac{2k}{K}\right)\binom{4K}{4k}\left(\binom{4k}{k}2^{-4k} - \binom{4(K-k)}{(K-k)}2^{-4(K-k)}\right) \quad > \quad \binom{4K}{K}2^{-4K}.$$

Or, equivalently:

$$\sum_{1\leq k<K/2}\left(1 - \frac{2k}{K}\right)\binom{4K}{4k}(F(k) - F(K-k)) \quad > \quad F(K), \tag{9}$$

[29] If $N \leq 2d$, then it follows that N_R is either N or $N/2$. Hence, we can infer directly that $Pr(X_n = R|\Theta = \theta) \geq 1/2$. For example, if $\Theta = 3/4$ and $N \leq 8$, since the number of landings in the red army zone must be a multiple of four, this number must be either four or eight. But, in either case, at least half of the total number of airdrops landed in the red army zone.

[30] In this sense, the assumption that the variable Θ contains all the probabilistically relevant information is enough to underwrite a judgement to the effect that Judy Benjamin is less likely than not to be in the red army zone. Arguably, this is the same sort of reasoning that underlies the '*a priori*' assessment of probabilities in the context of simple games of chance (see Vasudevan [2013]).

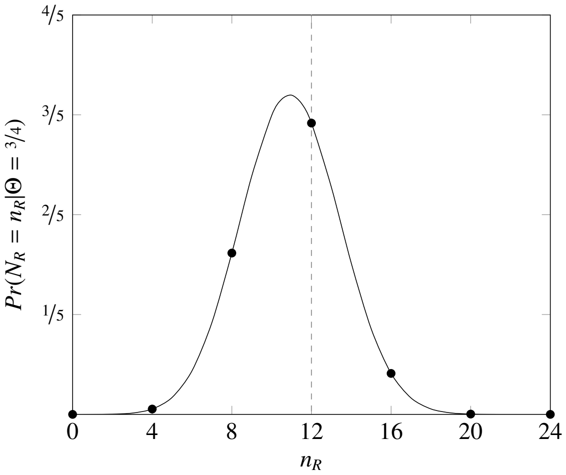


Figure 3: The probability that n_R of the variables $X_1, \ldots, X_{24}$ take values in the set $\{R_1, R_2\}$, given that $\Theta = 3/4$. Note that the distribution is skewed, obtaining its maximum for a value of n_R strictly less than 12.

where

$$F(n) = \binom{4n}{n} 2^{-4n}.$$

Since it can be shown that $F(n)$ is a decreasing function in n, the terms in the sum on the left-hand side of Equation (9) are all positive.[31] Hence, it suffices to establish the inequality to show that

$$\left(1 - \frac{2}{K}\right)\binom{4K}{4}(F(1) - F(K-1)) \quad > \quad F(K), \tag{10}$$

Since the left-hand side of this inequality is increasing in K and the right-hand side is decreasing in K, it is enough to show that the inequality holds for $K = 3$, that is:

$$\left(\frac{1}{3}\right)\binom{12}{4}(F(1) - F(2)) \quad > \quad F(3).$$

This can be confirmed directly.

Thus, we have shown that

$$Pr\,(X_n = R \,|\, \Theta = 3/4) < 1/2,$$

provided N is any multiple of 4 that is $\geq$ 12. Even without considering the details of the above derivation one can get some sense of why this should be so by observing the shape of the probability distribution described in Equation (7). Figure 3, for example, depicts this distribution for the case in which $N = 24$ and $\Theta = 3/4$. Note that the distribution peaks for a value of N_R that is strictly less than $N/2$.[32]

A second important point to note about Equation (8) is that, in the limit as $N \rightarrow \infty$, the quantity on the right-hand side of the equation converges uniformly in θ to $Pr^{\theta}_{\mathrm{ME}}(R)$ (see Figure 4).[33] Judy Benjamin would thus be led to adopt the maximum entropy estimate as the number

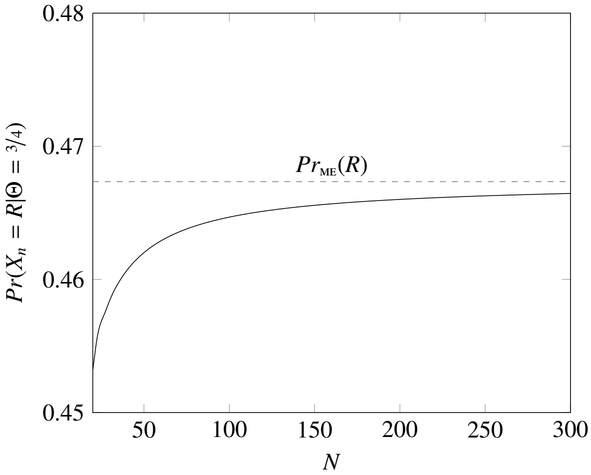


Figure 4: The probability, as a function of N, that any given X_n will take a value in the set $\{R_1, R_2\}$, given that $\Theta = 3/4$. As $N \rightarrow \infty$, this probability converges to the maximum entropy estimate $Pr_{\text{ME}}(R)$.

of observed airdrops grows increasingly large.

As the above analysis reveals, Judy Benjamin's apparently puzzling behaviour is entirely explicable provided we attribute to her the additional belief that the report from headquarters contains all the probabilistically relevant information relating to her situation. While it may not be obvious, this assumption imposes quite severe constraints on Judy Benjamin's reasoning, for it requires her to treat as equally probable all possible outcomes consistent with the report. As we have just seen, one such non-obvious constraint is that Judy Benjamin must accept that if $\Theta = 3/4$, then, on any given airdrop into the region, the probability of landing somewhere in the red army zone is strictly less than $1/2$.[34]

It is worth emphasizing that in relying on the assumption that the report from headquarters contains all the probabilistically relevant information related to her situation, Judy Benjamin is going well beyond the information contained in the report itself. In this sense, she is not acting conservatively. Rather, her behaviour reflects a certain charitable attitude towards her informant, which results in her adopting the view that, in formulating the report, the latter did not leave out any relevant information. Such charity on the part of Judy Benjamin is analogous to that which a student affords to an examiner in assuming that a certain problem on an exam is 'well posed'.[35] Even if the problem statement does not include any explicit claim to this

[31] That $F(n)$ is a decreasing function in n can be verified by appeal to the upper and lower bounds on the binomial coefficient given by Stănică ([2001], Theorem 2.6).

[32] Figure 3 is based on the standard gamma function interpolation of the factorial function, $\Gamma(n) = (n + 1)!$.

[33] This is, in fact, a corollary to Jaynes's concentration theorem, which implies that as $N \rightarrow \infty$, the possible outcomes satisfying $\Theta = 3/4$ grow increasingly concentrated about the maximum entropy values (see Jaynes [1983a]). In other words, an increasing proportion of these outcomes (approaching to 1) have an entropy which differs from the maximum entropy values by a decreasing amount (approaching to 0).

[34] This explanation generalizes to other cases of maximum entropy reasoning. Suppose, for example, that the four regions are assigned the scores $1, 2, 3$ and 4, respectively, with the middle scores 2 and 3 being assigned to the two regions in red army territory. Were Judy Benjamin to receive a report from headquarters informing her that the average score of airdrops into the area was some value other than 2.5, based on the principle of maximum entropy, she would assign a probability strictly less than $1/2$ to the event that she is somewhere in the red army zone (see Footnote 22 above). This inference can likewise be explained by attributing to Judy Benjamin the view that her informant's report of the average score contains all the probabilistically relevant information related to her situation.

[35] The phrase 'well-posed problem' is borrowed from Jaynes ([1973]). While Jaynes recognized the crucial role

effect, the pragmatics of exam-sitting require the student to extend charity to his examiner by presupposing that the problem, as stated, does not leave out any information that is relevant for determining its solution. In certain cases, this presupposition itself can impose quite substantial constraints on what can count as a correct solution to the problem.

We can therefore offer at least this much by way of a defence of Judy Benjamin. Suppose that Judy Benjamin were taking an exam and the following question were put to her:

> Given that $3/4$ of all red-zone landings land in the red army's headquarters region, what is the probability that any given airdrop will land somewhere in the red army zone?

In this case, she should not lose marks for giving an answer to this question that is strictly less than $1/2$. If she did, she could legitimately complain that her examiner had set her an ill-posed problem. Admittedly, this is a modest apologia for Judy Benjamin. For one thing, it is not at all clear that Judy Benjamin's situation, as portrayed in the original description of the scenario, is analogous to that of an examinee; nor is it obvious that her informant warrants the sort of epistemic charity she would have to extend to him in order to justify her conclusion that she is less likely than not to be in the red army zone. Whether or not this is so, will depend on how the finer details of the scenario are fleshed out.[36] More importantly, however, such a defence of Judy Benjamin clearly cannot be applied to justify the use of maximum entropy methods in theoretical contexts of inquiry. Nature, after all, is not an examiner who sets us problems to solve—we must raise the problems for ourselves. Consequently, if we are to base our probabilistic analyses on the assumption that a certain problem is well posed, our reasons for doing so must be grounded in theoretical knowledge and not mere pragmatic convention.

4 Conclusion: Entropy and Insufficient Reason

The Judy Benjamin problem represents a challenge for the principle of maximum entropy only insofar as this principle is viewed as a formalization of the conservative maxim to assume as little as possible within the constraints set by the available evidence. As we noted in the previous section, however, Judy Benjamin's seemingly odd behaviour can be rationalized provided we impute to her the non-conservative aim of exercising charity towards her informant by assuming that the report she receives from headquarters contains all the probabilistically relevant information related to her situation.

Thus, the moral of the Judy Benjamin problem is not that the principle of maximum entropy must be rejected, but rather that its justification must be reconceived. Maximum entropy reasoning should not be viewed as a prescription for how to remain maximally non-committal in the face of new evidence; it is rather a method for formalizing one's belief that a certain quantity contains all the relevant information needed to assess the relative probabilities of the possible outcomes of a given chance process. Or, to put it more succinctly, it is a method for formalizing one's belief that a certain problem in probability theory is well posed. Having

that assumptions of well-posedness play in maximum entropy reasoning, he generally held that the importance of these assumptions was limited to prior selection over a continuous parameter space. Moreover, Jaynes never seemed to view these assumptions of well-posedness as threatening to his conservative characterization of the maximum entropy distribution as maximally uncertain or minimally informative.

[36] Thus, for example, in van Fraassen's original description of the scenario, immediately after receiving the report from headquarters, Judy Benjamin's radio 'gives out' (van Fraassen [1981], p. 377). Now, if Judy Benjamin believes that her communication with headquarters was interrupted before the report was completed, she would obviously have no grounds for assuming that the part of the report that she did manage to receive contained all the probabilistically relevant information. Hence, in this case, she would have no grounds for applying the principle of maximum entropy.

made such a judgement, the choice to adopt the maximum entropy distribution follows from the further requirement that any two distinct outcomes that agree in all relevant respects must be assigned the same probability.

Note that this last requirement is just another way of formulating the old Laplacean principle of insufficient reason. Thus, contrary to Jaynes's view, the principle of maximum entropy, far from being a generalization of the principle of insufficient reason, turns out to be a direct corollary of it. This, of course, will come as no surprise to those familiar with the applications of maximum entropy reasoning that arise in statistical mechanics. In these contexts, the derivation of the maximum entropy distribution from conditionally uniform probability measures is fully explicit, even if the judgements of relevance that underwrite these measures are couched in the technical jargon of 'partition functions' and 'microcanonical ensembles'.[37] Indeed, on several occasions, Jaynes himself made note of the essential role that assumptions of probabilistic relevance play in maximum entropy reasoning.[38] Yet, for some reason, Jaynes never took such assumptions to threaten his characterization of the maximum entropy distribution as minimally informative or maximally non-committal. In fact, he rarely discussed the grounds for such assumptions, treating them, in most cases, as mere pragmatic conventions that presuppose no substantive theoretical knowledge.[39]

At this point, it would be easy to criticize Jaynes for being philosophically careless. In his overeagerness to portray the principle of maximum entropy as an *a priori* precept of rationality, he simply neglected the background theoretical assumptions required for its legitimate application. Such a targeted critique, however, overlooks the more fundamental philosophical challenges that underwrite Jaynes's confusion. The real difficulty stems from the fact that we do not have a satisfactory account of how our judgements of probabilistic relevance are grounded in objective knowledge of the world. As a result, we are mistakenly led to interpret these judgements in subjective terms.

This philosophical confusion predates Jaynes and is manifest even in the writings of the classical Bayesians. As a result of their commitment to a deterministic worldview, the classical Bayesians had difficulty seeing how the judgements of 'equipossibility' underwriting their appeals to the principle of insufficient reason could have objective purport. Consequently, they were led to adopt a subjective interpretation of these judgements according to which two distinct outcomes are equally possible just in case an agent is in a state of total ignorance as to which of these outcomes is more likely to occur. While this may seem like a modest conceptual

[37] See (Martin-Löf [1974]) for an exposition of the statistical mechanical terminology along these lines in terms of repetitive structures and sufficient statistics.

[38] For example, in responding to an objector, who asked rhetorically whether there is anything in the physics of throwing dice to suggest the plausibility of the maximum entropy distribution, Jaynes ([1978], pp. 266–7)., after describing the very detailed physical experiment he has in mind, observed:

> Success in using [the principle of maximum entropy] does not require that we take into account all dynamical details; it is enough if we can recognize whether by common sense analysis or by inspection of the data, *what are the systematic influences at work, that represent the 'physical constraints?'* If by any means we can recognize these, maximum entropy then takes over and supplies the rest of the solution.

As this passage clearly indicates, the constraints to which the principle of maximum entropy is meant to be applied must summarize all the 'physical constraints' in operation in a given experiment. In other words, they must exhaust all the probabilistically relevant information. Note, however, Jaynes's insistence that the physical constraints can somehow be discerned through 'common sense analysis' or 'inspection of the data', without any appeal to background theoretical considerations.

[39] For example, Jaynes ([1973], p. 146) describes the presupposition that a given problem is well posed as a 'matter of courtesy' (see also Jaynes [1968], p. 239).

shift, in point of fact, this subjective interpretation of the conditions under which the principle of insufficient reason can be applied gives rise to a wholesale inversion in the order of our probabilistic knowledge, and has been the underlying source of a great deal of confusion.

Consider, for example, the case of casting a fair die. On a subjective reading, our assignment of equal probabilities to each of the six possible outcomes of the cast is reflective of a state of total ignorance as to the relative probabilities with which these outcomes will occur. In truth, however, the situation is exactly the opposite. What justifies us in assigning an equal probability to the possible outcomes of the cast is not the fact that we lack any information describing the statistical behaviour of the die, but rather our belief that, even absent any such information, the problem of assessing the relative probabilities of the possible outcomes of the cast is well posed. Of course, the less information one takes to be needed in order for a problem to be well posed, the more information one must already possess—*ex nihilo nihil* is, after all, as much a principle of epistemology as of metaphysics. Thus, in the case of simple games of chance, our appeal to the principle of insufficient reason is not predicated on total ignorance, but instead on substantial theoretical knowledge encoded in our intuitive capacity to recognize the symmetries exhibited by the chance processes employed in such games.[40]

The same point applies to the principle of maximum entropy. If the constraints to which the principle is applied are taken to reflect an agent's subjective state of knowledge, then weaker constraints (and more uniform distributions) are indicative of a greater degree of ignorance as to the outcome. But, if these constraints are instead viewed as specifying the information needed in order for a problem to be well posed, then to apply the principle of maximum entropy to weaker constraints and thus obtain more uniform distributions requires more—not less—theoretical knowledge.[41]

Jaynes, of course, was right to observe that the analysis of games of chance, which was the main preoccupation of the classical theorists of probability, represents only a small part of the general theory of probabilistic reasoning. This, however, is not because total ignorance is rare and that we almost always possess at least some relevant information. It is rather owing to the fact that we have a far more complete theoretical understanding of the highly refined and carefully engineered chance processes employed in games of chance than we do of almost all other stochastic phenomena that we encounter outside of the casino.

In Jaynes's view, the philosophical problem bequeathed to us by the classical Bayesians was to develop a theory of probabilistic inference that could extend beyond combinatorial reasoning based on the principle of insufficient reason. Ironically, the principle of maximum entropy provides evidence for the claim that such combinatorial reasoning may, in the end, be all that we need. For, maximum entropy reasoning shows us how the principle of insufficient reason can structure our probabilistic analyses even in cases where our theoretical knowledge of a chance process is less than complete.[42] Thus, the only real philosophical challenge that remains

[40] This is reflected in the obvious, but sometimes overlooked, fact that the randomizing processes employed in simple games of chance are not arbitrarily chosen, but are rather refined products of human engineering that that have been carefully designed to ensure that they exhibit the requisite symmetries (see David [1962], Chapter 1).

[41] This apparent inversion is highlighted by the objection to the principle of maximum entropy raised at (Seidenfeld [1979], pp. 429–34).

[42] A similar point can be made with respect to de Finetti's theorem for exchangeable variables (which is, in fact, just a particular instance of maximum entropy reasoning). De Finetti's own subjectivist inclinations led him to view this thoerem as a way to make sense of frequentist reasoning without positing objective probabilities (see de Finetti [1980], Chapters 4, 5). In fact, de Finetti's theorem is better understood as indicating how frequentist reasoning can be justified on the grounds of the principle of insufficient reason applied to the belief that the binomial statistic contains all the information relevant to assessing the relative probabilities of the outcomes of a discrete-time stochastic process that takes only two values.

is to explain how the principle of insufficient reason, and the judgements of equipossibility on which it is based, can be given an objective interpretation. Or, to put it in Jaynes's terms, the challenge is to explain how our objective knowledge of the world gets encoded in the form of our capacity to judge whether a given problem is well posed.

Anubav Vasudevan
Department of Philosophy
University of Chicago
Chicago, USA
anubav@uchicago.edu

References

Bernoulli, J. [2006]: *The Art of Conjecturing, Together with Letter to a Friend on Sets in Court Tennis*, Johns Hopkins University Press.

Clarke, B. S. and Barron, A. R. [1994]: 'Jeffreys' Prior Is Asymptotically Least Favorable under Entropy Risk', *Journal of Statistical Planning and Inference*, **41**, pp. 37–60.

Cox, R. T. [2001]: *Algebra of Probable Inference*, Johns Hopkins University Press.

Csiszar, I. [1991]: 'Why Least Squares and Maximum Entropy? An Axiomatic Approach to Inference for Linear Inverse Problems', *The Annals of Statistics*, **19**, pp. 2032–66.

David, F. [1962]: *Games, Gods, and Gambling: A History of Probability and Statistical Ideas*, Dover Publications.

de Finetti, B. [1980]: 'Foresight: Its Logical Laws, Its Subjective Sources', in H. Kyburg and H. Smokler (*eds*), *Studies in Subjective Probabilities*, Robert E. Krieger, pp. 53–18.

Dias, P. M. and Shimony, A. [1981]: 'A Critique of Jaynes' Maximum Entropy Principle', *Advances in Applied Mathematics*, **2**, pp. 172–211.

Douven, I. and Romeijn, J.-W. [2011]: 'A New Resolution of the Judy Benjamin Problem', *Mind*, **120**, p. 637.

Feller, W. [1968]: *An Introduction to Probability Theory and Its Applications*, Volume 1, Wiley.

Friedman, K. and Shimony, A. [1971]: 'Jaynes's Maximum Entropy Prescription and Probability Theory', *Journal of Statistical Physics*, **3**, pp. 381–4.

Gaifman, H. and Vasudevan, A. [2012]: 'Deceptive Updating and Minimal Information Methods', *Synthese*, **187**, pp. 147–78.

Goldstein, S. [2001]: 'Boltzmann's Approach to Statistical Mechanics', in J. Bricmont, G. Ghirardi, D. Dürr, F. Petruccione, M. C. Galavotti and N. Zanghi (*eds*), *Chance in Physics: Foundations and Perspectives*, Heidelberg: Springer, pp. 39–54.

Grove, A. and Halpern, J. Y. [1997]: 'Probability Update: Conditioning vs. Cross-entropy', in D. Geiger and P. Shenoy (*e*ds),*Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, San Fransisco, CA: Morgan Kaufmann, pp. 173–82.

Heath, D. and Sudderth, W. [1976]: ‘De Finetti’s Theorem on Exchangeable Variables’, *The American Statistician*, **30**, pp. 188–9.

Huisman, L. [2014]: ‘On Indeterminate Updating of Credences’, *Philosophy of Science*, **81**, pp. 537–57.

Jaynes, E. T. [1957a]: ‘Information Theory and Statistical Mechanics, I’, *Physical Review*, **106**, pp. 620–30.

Jaynes, E. T. [1957b]: ‘Information Theory and Statistical Mechanics, II’, *Physical Review*, **108**, pp. 171–90.

Jaynes, E. T. [1967]: ‘Foundations of Probability Theory and Statistical Mechanics’, in M. Bunge (*ed.*), *Delaware Seminar in the Foundations of Physics*, Springer.

Jaynes, E. T. [1968]: ‘Prior Probabilities’, *IEEE Transactions on Systems Science and Cybernetics*, **4**, pp. 227–41.

Jaynes, E. T. [1973]: ‘The Well-Posed Problem’, *Foundations of Physics*, **3**, pp. 477–93.

Jaynes, E. T. [1983a]: ‘Concentration of Distributions at Entropy Maxima’, in R. D. Rosenkrantz (*ed.*), *E. T. Jaynes: Papers on Probability, Statistics, and Statistical Physics*, D. Reidel, pp. 315–36.

Jaynes, E. T. [1983b]: ‘Where Do We Stand on Maximum Entropy?’, in R. D. Rosenkrantz (*ed.*), *E. T. Jaynes: Papers on Probability, Statistics, and Statistical Physics*, D. Reidel, pp. 210–314.

Jeffreys, H. [1961]: *Theory of Probability*, Oxford: Oxford University Press.

Kass, R. E. and Wasserman, L. [1996]: ‘The Selection of Prior Distributions by Formal Rules’, *Journal of the American Statistical Association*, **91**, pp. 1343–70.

Keynes, J. [1921]: *A Treatise on Probability*, Macmillan.

Martin-Löf, P. [1974]: ‘Repetitive Structures and the Relation between Canonical and Microcanonical Distributions in Statistics and Statistical Mechanics’, in O. Barndorff-Nielsen, P. Blæsild and G. Schou (*eds*), *Proceedings of Conference on Foundational Questions in Statistical Inference, Aarhus, 1973*, Aarhus: University of Arrhus pp. 271–94

Paris, J. and Vencovská, A. [1990]: ‘A Note on the Inevitability of Maximum Entropy’, *International Journal of Approximate Reasoning*, **4**, pp. 183–223.

Pierce, J. R. [1980]: *An Introduction to Information Theory*, Dover.

Seidenfeld, T. [1979]: ‘Why I Am Not an Objective Bayesian; Some Reflections Prompted by Rosenkrantz’, *Theory and Decision*, **11**, pp. 413–40.

Seidenfeld, T. [1987]: ‘Entropy and Uncertainty’, in I. B. MacNeill, G. J. Umphrey, M. Safiul Haq, W. L. Harper and S. B. Provost (*eds*), *Advances in the Statistical Sciences: Foundations of Statistical Inference*, Volume 2, Springer, pp. 259–87.

Shannon, C. E. [1948]: ‘A Mathematical Theory of Communication’, *Bell System Technical Journal*, **27**, pp. 379–423.

Shimony, A. [1973]: ‘Comment on the Interpretation of Inductive Probabilities’, *Journal of Statistical Physics*, **9**, pp. 187–91.

Shore, J. E. and Johnson, R. W. [1980]: ‘Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy’, *IEEE Transactions on Information Theory*, pp. 26–37.

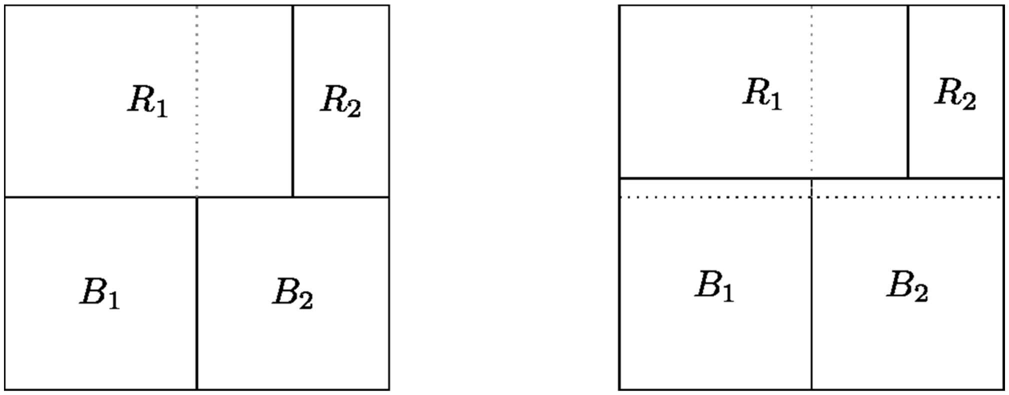
Stănică, P. [2001]: ‘Good Lower and Upper Bounds on Binomial Coefficients’, *Journal of Inequalities in Pure and Applied Mathematics*, **2**, pp. 1–5.

van Fraassen, B. C. [1981]: ‘A Problem for Relative Information Minimizers in Probability Kinematics’, *British Journal for the Philosophy of Science*, **32**, pp. 375–9.

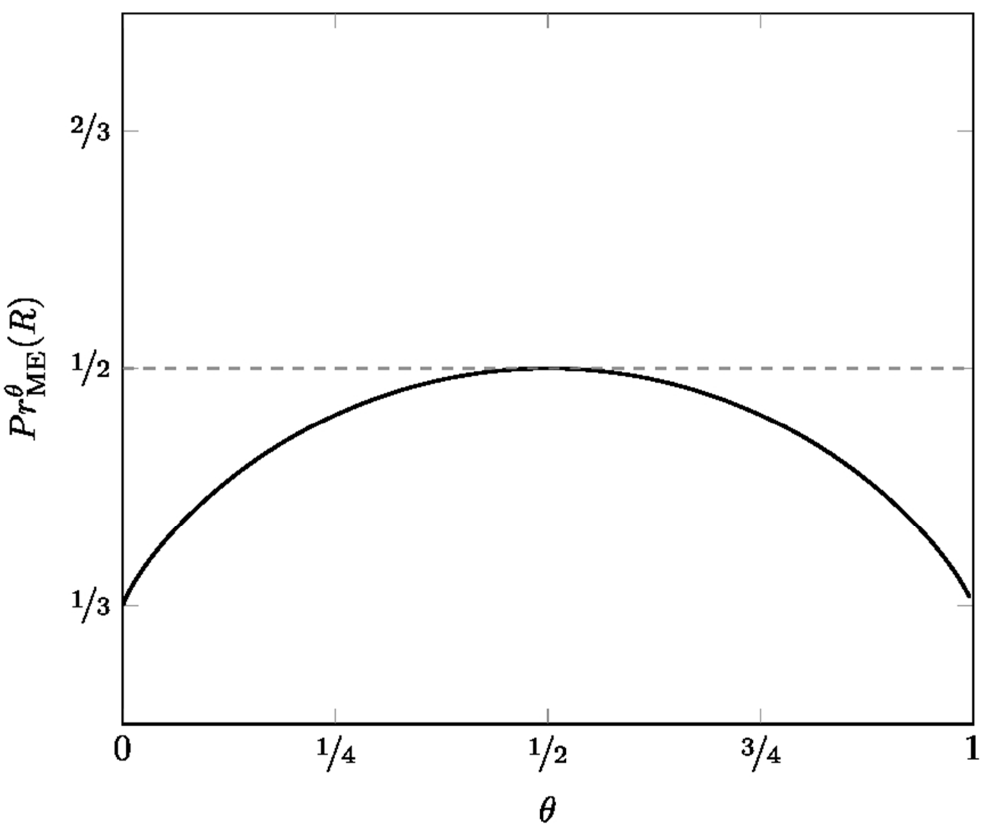
van Fraassen, B. C., Hughes, R. I. G. and Harman, G. [1986]: ‘A Problem for Relative Information Minimizers, Continued’, *British Journal for the Philosophy of Science*, **37**, pp. 453–63.

Vasudevan, A. [2013]: ‘On the *a priori* and *a posteriori* Assessment of Probabilities’, *Journal of Applied Logic*, **11**, pp. 440–51.

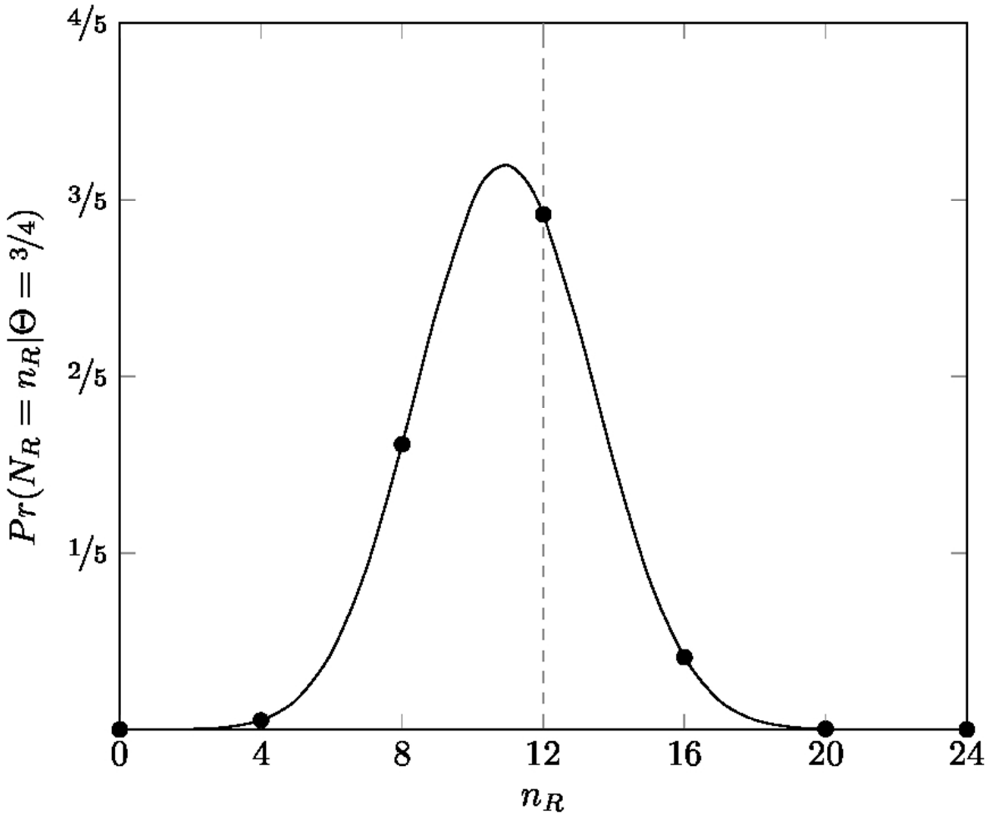
Williamson, J. [2010]: *In Defence of Objective Bayesianism*, Oxford: Oxford University Press.



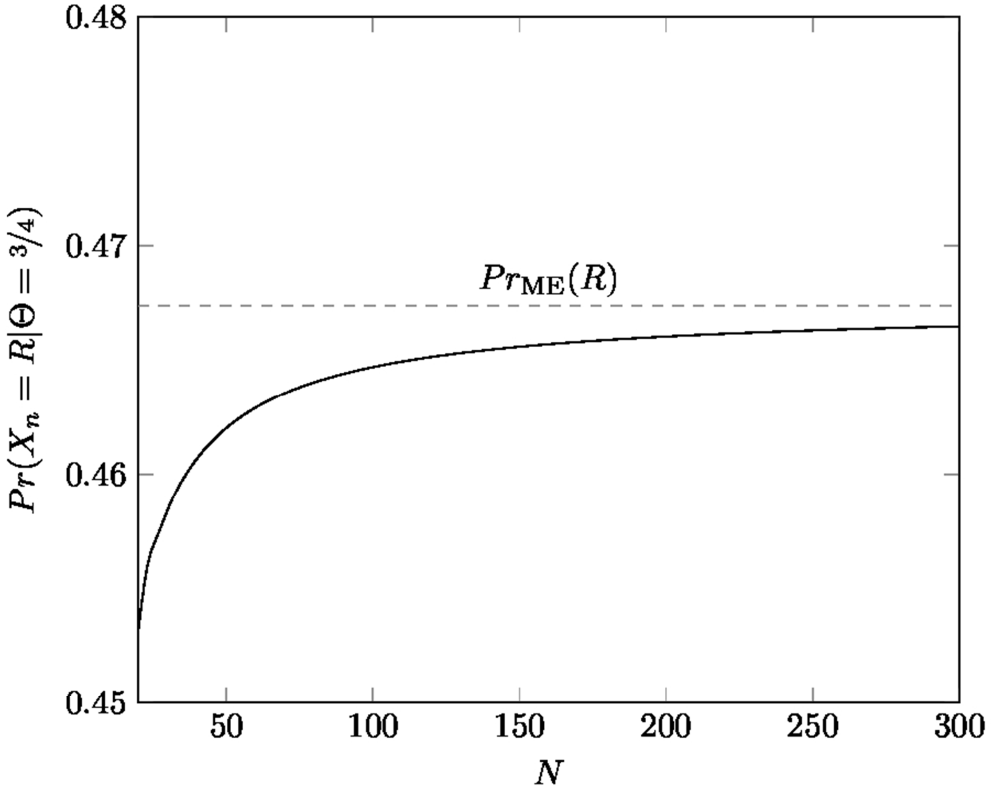
93x36mm (300 x 300 DPI)



74x60mm (300 x 300 DPI)



74x61mm (300 x 300 DPI)



76x61mm (300 x 300 DPI)