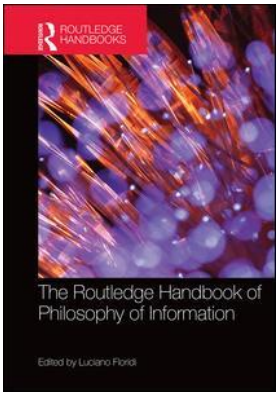


# The Routledge Handbook of Philosophy of Information

Luciano Floridi

## Probability and information

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9781315757544.ch02
Peter Milne
**Published online on: 23 Jun 2016**

# 2
# PROBABILITY AND INFORMATION

*Peter Milne*

## Introduction

Serious thought about probability, both as degree of belief in the light of evidence and as reflecting the tendency to produce stable relative proportions of occurrence upon repetition (as with the ratio of heads to tails when a coin is tossed repeatedly), emerged in the middle of the seventeenth century (see Hacking 2006). While probability and the related notions of likelihood and chance are nowadays in part everyday notions, they have also been regimented or codified in the formal, mathematical theory of probability. This formal theory admits various interpretations, some but not all of which draw on the everyday notions. Here I shall sketch connections between information and some interpretations of the formal theory. I shall begin by introducing the bare bones of the mathematical theory, sufficient to the demands of this chapter.

## Probability: the mathematical theory

Probabilities are assigned to events or, more exactly, *distributed over* a *family* or *field of events*. This field has the structure of a *Boolean algebra*; that is, it contains: (*i*) the certain event $S$ which is sure to occur, and the impossible event $\varnothing$ which cannot occur; (*ii*) if it contains the event $e$, it contains the complementary event not-$e$ which occurs just if $e$ does not; (*iii*) if it contains the events $e$ and $f$, it contains the event $e \,\&\, f$ of their joint occurrence and the event $e \vee f$ that occurs when at least one of $e$ and $f$ occurs ('$\vee$' from the Latin word 'vel' meaning 'or'). An assignment *prob* of numerical values to the members of the field of events is a *probability distribution* just in case it satisfies these *principles* or *axioms*:

1 for every event $e$, $0 \leq prob(e) \leq 1$;
2 $prob(S) = 1$; $prob(\varnothing) = 0$;
3 if the joint occurrence of $e$ and $f$ is impossible, i.e., if $e \,\&\, f = \varnothing$, $prob(e \vee f) = prob(e) + prob(f)$.

From these axioms it follows that $prob(e) + prob(\text{not-}e) = 1$, for $e$ and not-$e$ are jointly impossible and the event '$e \vee$ not-$e$' is certain.

To this we must add the definition of *conditional probability*. *prob*($f \mid e$), read "the probability of *f* given *e*", is defined as follows when *prob*($e$) > 0 (and is undefined otherwise):

$$prob\left(f \middle| e\right) = \frac{prob(f \,\&\, e)}{prob(e)}$$

As the joint occurrence of *e* with itself is just the event of *e*'s occurrence, *prob*($e \mid e$) = 1; since the joint occurrence of *e* and not-*e* is impossible, *prob*(not-$e \mid e$) = 0.

Some authors take the notion of conditional probability as basic. For each event *e*, they take the function *prob*( . | *e*) to assign numerical values to members of the field of events in accordance with the principles (1)–(3) above and add the extra constraint:

for any events $e$, $f$ and $g$, $prob(f \,\&\, g \mid e) = prob(g \mid e \,\&\, f) \times prob(f \mid e)$.

This makes sense even when $prob(f \mid e) = 0$, for $prob(f \,\&\, g \mid e) = 0$ too in this case.

## Information and probability as subjective degree of belief

There is a very straightforward connection between probability and information: the more likely you think it is that an event will occur, the more strongly you expect it to occur, the less surprised you are when it does occur, to the point that if you are certain it will occur, its occurrence is "no news to anybody." The more convinced you are that it will occur, the less you feel you have learned when informed that it has occurred. A newspaper that reported only the obvious, platitudinous, and well known would be a newspaper in name only, it would contain no news.

There are immediately a number of things that can be said about this particular linking of probability and information. One is that the conception of *information* involved here is that of "news value" or "surprise value"; another is that, given how I have set it up, it involves an individual's evaluations of what is likely and to what extent. To speak very loosely, one's beliefs constitute one's map of how things are; like ancient maps of the world, it contains *terrae incognitae* where various possibilities come to mind but one is not certain which is the case; however, some are more likely, maybe much more likely, than others. To be (newly) informed that such-and-such is the case is to fill in some chunk of *terra incognita* in one's map – and to wipe out or close down some of the possibilities previously entertained. To speak *very* loosely, the larger the chunk of *terra incognita* filled in, the larger the swathe of possibilities closed down, the more you have learned, the more information you have gained – and probability, if it is anything, is a measure of possibilities. Thus we are led to the idea that probability and information go in opposite directions: the more probable, the less informative, and *vice versa*. And thus we are led to the idea that information as news value or surprise value should be measured by some decreasing function of probability. Since what is certain affords no surprise, we want *inf*($e$) = 0 when *prob*($e$) = 1. One very simple measure meeting this constraint is *inf*($e$) = 1 – *prob*($e$).

Digging deeper, one might hold that when events *e* and *f* are uncorrelated, the information that one gains when one learns that both have occurred is the sum of what one learns from learning each has occurred since, being uncorrelated, neither bears on the other. This quite natural thought gives us the constraint

$$inf(e \,\&\, f) = inf(e) + inf(f) \text{ when } prob(e \,\&\, f) = prob(e) \times prob(f),$$

*prob*($e$ & $f$) = *prob*($e$) × *prob*($f$) being the probabilist's way of capturing lack of correlation.[1] For this to hold generally, we *must* have *inf*($e$) = –log(*prob*($e$)) (where the base of the logarithms may be chosen arbitrarily).[2] Following a path laid out by the statistician and philosopher

of science I. J. Good, we have arrived in a very straightforward way at one very common probability-based measure of information (Good 1950: 74–5).

We should look a little more closely at this. First, we started out from an individual's estimation of what is likely, what unlikely, and to what extent – from what, in the jargon, are known as *subjective probabilities*, *credences*, or *degrees of belief*. Much has been written on why a rational individual's degrees of belief ought to satisfy the standard mathematical framework of probability theory. Here we shall take for granted that they do. (Items in the further reading section present arguments for why this should be so.) What concerns us here is that different individuals may give different estimates of how likely an event is: our probability-based measures of information will inherit this subjectivity from degrees of belief. Moreover, an individual's estimate of how likely an event is may change over time, more particularly, with what the individual learns over time. Thus what we have here is a conception of information, and a measure to go with it, that may vary from individual to individual, and, for a single individual, may vary as what the individual takes herself to know changes over time.

The mere addition of new beliefs consistent with what one fully believed previously, i.e., with that to which one previously assigned maximum degree of belief, is most straightforwardly dealt with under the procedure known as *Bayesian conditionalization* (application of *Bayes' Rule*); much harder to model formally is the process of adding information that conflicts with what one previously believes – here belief revision theory tells part of the story but how to marry it with subjective probability is hardly a settled matter. *Subjective Bayesianism* adds to the subjective interpretation of probability as degree of belief updating of degree of belief by *Bayes' Rule*: upon learning that $e$ and nothing more, the rational individual revises her degrees of belief according to the schema

$$prob_{new}(f) = prob_{old}(f|e),$$

provided that $prob_{old}(e) \neq 0$. $prob_{old}$ is commonly called the prior probability, $prob_{new}$ the posterior.

I have spoken of individuals' estimates of how likely events are. Mathematical probability theory assigns probabilities to events and gives to the field of events over which a probability distribution distributes probabilities in the structure of a Boolean algebra. Philosophers are most likely to think of propositions as the sort of thing one believes: if one believes that Dushanbe is the capital of Tajikistan then *that Dushanbe is the capital of Tajikistan* is the proposition believed. Degrees of belief too, then, are degrees of belief in propositions and, almost invariably, subjective probabilities are assigned to propositions, not events. The difference, however, is small, for, on the assumption that a rational individual's degrees of belief are governed by classical propositional logic and that she assigns the same probability to logically equivalent propositions, we can recover the Boolean algebraic structure of the domain of that to which probabilities are assigned.

## Information and "logical probability"

All this inter- and intra-individual variability may lead one to think we are missing something important in the notion of information: we might think that how informative a message is has something to do with the content of the message, not how surprising (or not) its recipient finds it. In the 1950s this idea was tackled, still in probabilistic terms, in Yehoshua Bar-Hillel and Rudolf Carnap's notions of *content* and *semantic information* (Bar-Hillel and Carnap 1953) and Karl Popper's notion of *content* (Popper 1957, 1959, Appendix *IX). Both parties take probability as basic; both parties recognize 1 – *prob* and –log *prob* as possible measures. What is different here is the conception of probability. Here the probabilities in question

are measures of logical strength, running from zero for logical contradictions such as '*e* & not-*e*' to 1 for logical tautologies, propositions such as '*e* ∨ not-*e*' which cannot but be true. Carnap spent much of the later part of his career attempting to spell out the details of how to assign so-called *logical probabilities* to the sentences of formal languages in such a way as to accommodate a "logical" account of inductive reasoning. As the project progressed, more parameters entered the system so that what looks more and more like the variability from individual to individual of subjective probability becomes a part of the theory of supposedly logical probability. The widespread – but not universal – consensus among philosophers is that the project failed; and the current popularity of subjective probability (and Subjective Bayesianism) in the philosophical literature is in no small part due to this perceived failure.

## Information and the classical conception of probability

Let's back-track a little. The idea of a logical conception of probability assigned to propositions first emerged – one source is Ludwig Wittgenstein's *Tractatus Logico-Philosophicus* – as a formal analogue of the *classical conception of probability* according to which the probability of an outcome just is the ratio of the number of cases favorable to the outcome to the number of all possible cases. One needs here a specification of the possible cases but in applications this is often just obvious – e.g., the six sides of a die or the 52 cards in a standard deck. (The aces of the four suits are the cases favorable to being dealt an ace so the probability of being dealt an ace is $4 \div 52 = \frac{1}{13}$.) The classical conception works wonderfully well for games of chance – dice, cards, roulette, for example – but is rather less useful if one wishes to bet on horses. For horse-racing one needs more information than just the number of runners: one needs to know the horses' recent form, the state of the ground, and something about the jockeys up that day – and maybe training regimes, the likelihood of each horse being doped, and goodness knows what else. It's all much simpler in the cases of dice, cards, and roulette wheels, or so it seems. If you have no special information to go on – and mostly you don't – you have no reason to expect one face of the die uppermost rather than another, one card rather than another, the ball to end up in one slot on the roulette wheel rather than another. One has one's possible cases to each of which – as the classical conception requires – one assigns equal probabilities ($\frac{1}{6}$, $\frac{1}{52}$, $\frac{1}{37}$ in Europe, $\frac{1}{38}$ in North America). In the absence of information pointing to one outcome rather than another, assigning equal probabilities to the basic possible cases seems the right, the reasonable, the rational thing to do. By assigning equal probabilities, one isn't building in information one hasn't got. And when you get the information as to which outcome occurred, each of the possible outcomes is equally – and maximally – informative. In advance, one expects to gain the same amount of information, whichever outcome occurs. Hold that thought!

## Entropy

Now, take the measure of information –log *prob*. A *partition* is a set of mutually exclusive and jointly exhaustive events, such as the classical conceptions "possible cases": exactly one has to occur. Given a partition $\{e_1, e_2, \ldots, e_n\}$ of $n$ events, let $X(e_i)$ be a quantity associated with each event, possibly but not necessarily varying from event to event in the partition. The (*mathematical*) *expectation* or *expected value* or *mean* of the quantity $X$ with respect to the partition $\{e_1, e_2, \ldots, e_n\}$ is the sum

$$\Sigma_i^n =_1 X(e_i) \times prob(e_i).$$

Expectations are a bit like averages. To force through the analogy, think of $prob(e_i)$ as the proportion of cases yielding the value $X(e_i)$. The notion of expected value must be treated with care for the expected value of a quantity may not be a realisable value of that quantity; for example, with a fair die, one for which the probability of each face falling uppermost is $\frac{1}{6}$, the expected value of the number of spots on the uppermost face is 3.5 but no face has three and a half spots painted on it.

In the particular case in which the quantity of interest is information and we measure it by –log *prob*, the sum in question is

$$-\Sigma_i^n =_1 (prob(e_i) \times \log prob(e_i)) \ .^3$$

Due to a formal similarity with the physical quantity of the same name, this is called the *entropy* of the distribution *prob* with respect to the partition $\{e_1, e_2, \ldots, e_n\}$. Now, assigning different probabilities to the members of the partition may yield different values for the entropy and it is a mathematical fact of no little interest in the present context that the sum takes its maximum value when we assign the same probability to each of $e_1, e_2, \ldots, e_n$. We maximize entropy/expected information by assigning equal probabilities (as the classical conception says we ought).

## Objective Bayesianism and the principle of maximum entropy

We'd like to say that we maximize the information we expect to get on learning the actual outcome by adopting the classical conception's uniform distribution, the assignment of probabilities that recommends itself by not "building in information we haven't got." Unfortunately, on closer inspection this thought may appear no more than a pun on different uses of the words "expect" and "expectation". It would take considerably more space than I have at my disposal here to defend the claim that it is not. Suffice it to say now that we have just encountered the basic result of the classical conception's closest modern descendant, *Objective Bayesianism*.

Objective Bayesianism, prominently championed by the physicist E. T. Jaynes among others, enjoins – at least in some of its guises – the rational individual to assign as degrees of belief that prior distribution of probabilities that maximizes entropy. (It merits the epithet "Bayesianism" because it accepts Bayes' Rule for updating degrees of belief.) Two comments are called for, one technical, one conceptual. First, the technical. There is no straightforward extension from "discrete" probabilities assigned to the members of a finite partition to continuous probability distributions, distributions such as the normal distribution, although there is a widely accepted work-around: the (relative) entropy of the assignment $prob_2$ to the members of the partition $\{e_1, e_2, \ldots, e_n\}$ relative to the "reference" distribution $prob_1$ is given by the sum

$$\Sigma_i^n =_1 prob_2(e_i) \times \log \frac{prob_2(e_i)}{prob_1(e_i)} \ ;$$

this notion readily extends to the continuous case.[4] Second, the conceptual. Unlike the classical conception which mandates a uniform distribution, one can apply the rubric of maximizing entropy subject to constraints which do "point to one outcome rather than another"; from a long series of tosses of a die, one may learn that it is biased, giving a mean number of spots of 4.5, not the 3.5 obtained from the uniform distribution; one can maximize entropy subject to the constraint that the expectation be 4.5 and obtain a distribution skewed

in favor of the faces with larger numbers of spots. One uses the information one has and still doesn't "build in information one hasn't got."

In a rather beautiful confluence of ideas, we have that the assignment of probabilities that minimizes (relative) entropy relative to the reference distribution $prob_{old}$ subject to the constraint that $prob(e) = 1$ is the distribution $prob_{new}$ obtained by Bayes' Rule (Williams 1980: 134–135). Here *minimization*, rather than maximization, is appropriate since we seek to make the least change consistent with the constraint.

## Entropy in Shannon's approach

This far we have focused on probability as degree of belief, whether subjective or in some way objectively prescribed, taken information, or, perhaps better, informativeness to be a property of events, and arrived at entropy as the expected value of the latter quantity. While entirely natural given the Bayesian framework widely adopted in contemporary philosophy of science and formal epistemology, this is not at all how entropy entered into mathematical information theory. In Shannon's theory we deal with *statistical probabilities* – proportions of occurrence of signal items in a large sample of messages – and, quite generally, want a measure of the *uncertainty* associated with the assignment of probabilities to the members of a partition. Shannon lays down some desiderata that the measure should meet and proves that entropy, as defined above, is the unique measure meeting them (Shannon 1948, Appendix 2; Khinchin 1953). The important point here is that uncertainty is a property of the probability distribution as a whole. We do not start out from a quantity assigned to each of the members of the partition. This is not at all to deny that Shannon measures the information associated with a type of signal item $e$ by $-\log prob(e)$ where $prob(e)$ is $e$'s statitistical probability of occurrence.

## A change of direction: from information to probability

All of the above has taken probability as basic and has measured information or entropy/uncertainty in terms of it. Lastly, and very briefly, we look at an approach that reverses that direction. I. J. Good suggested "the possibility of deriving the axioms of probability from the concept of information instead of the other way round" and took some steps towards doing so (Good 1966); the project is developed further in (Milne 2012). Milne considers the amount of information $e$ adds to $f$ and lays down some intuitive constraints. $inf(e, e \,\&\, f) = 0$ and $inf(e, g) \leq inf(e \,\&\, f, g)$ are obvious ones. One, proposed by Good, which does a lot of work is this:

> $inf(e \,\&\, f, g)$ is determined by $inf(e, g)$ and $inf(f, e \,\&\, g)$.

The information $e$ and $f$ jointly add to $g$ is fixed by the amount of information $e$ adds to $g$ and the amount $f$ adds over and above that once $e$ has been "taken on board."

Milne distinguishes two conceptions of information: one, a "novelty value" conception, adds the constraint $inf(e, f \,\&\, g) \leq inf(e, f)$ for $e$ can't be more novel with respect to a larger corpus of information than with respect to a smaller one; the other views $inf(e, f)$ as a measure of the proportion of possibilities left open by $f$ that are closed down by $e$ and holds that $inf(\text{not-}e, f)$ is determined by $inf(e, f)$ for possibilities not closed down by $e$ are closed down by not-$e$ and *vice versa*.[5] The former leads to a measure that rescales to a non-standard, probability-like function similar to those found in (Morgan and Mares 1995); the other leads to a measure that rescales as a conditional probability distribution along the lines laid out by R. T. Cox (Cox 1946; Cox 1961, Chapter 1).

For more on Bayesianism, see Chaper 16 of this handbook. For more on Shannon's work and the mathematical theory of information, see Chapter 4. For more on conceptions of semantic information, see Chapter 6.

## Notes

1 $prob(e\ \&\ f) - (prob(e) \times prob(f))$ and $\dfrac{prob(e\ \&\ f)}{prob(e) \times prob(f)}$ have both been suggested as quantitative measures of correlation, one taking the value 0, the other 1, when $e$ and $f$ are uncorrelated. Although it may not be obvious, the first of these is equivalent to $(prob(e\ \&\ f) \times prob(\text{not-}e\ \&\ \text{not-}f)) - (prob(e\ \&\ \text{not-}f) \times prob(\text{not-}e\ \&\ f))$ and, under the name *the odds ratio*, the quantity $\dfrac{prob(e\ \&\ f) \times prob(\text{not-}e\ \&\ \text{not-}f)}{prob(e\ \&\ \text{not-}f) \times prob(\text{not-e}\ \&\ f)}$ is a widely used measure of correlation in medical statistics.

2 For $x > 0$, the logarithm to base 2 of $x$, written $\log_2 x$, is that number $y$ such that $x = 2^y$; the logarithm to base 10 of $x$, written $\log_{10} x$, is that number $z$ such that $x = 10^z$; $y$ and $z$ are related by the conditions $y = z \times \log_2 10$ and $z = y \times \log_{10} 2$. ($\log_2 10 \approx 3.322$; $\log_{10} 2 \approx 0.301$).

Below we shall write $-\log prob(e)$ rather than $-\log(prob(e))$.

3 We stipulate that $prob(e_i) \times \log prob(e_i) = 0$ when $prob(e_i) = 0$.

4 The relative entropy is also called the *Kullback-Leibler divergence*. 'Divergence' because *in some respects* this quantity functions like a measure of how far apart the two distributions are – it is minimized when $prob_2$ is identical to $prob_1$.

5 We are thinking here of maximally specific possibilities so that in any possibility either $e$ obtains or not-$e$ obtains.

## Further Reading

Childers, Timothy (2013) *Philosophy and Probability*, Oxford: Oxford University Press. A critical (and approachable) guide to conceptions of probability including Objective Bayesianism and the principle of maximum entropy.

Cox, R. T. (1961) *The Algebra of Probable Inference*, Baltimore, MD: Johns Hopkins University Press. A stimulating but idiosyncratic approach to the foundations of probability and to information-theoretic entropy.

Gigerenzer, G. *et al.* (1989) *The Empire of Chance: How Probability Changed Science and Everyday Life*, Cambridge: Cambridge University Press. A survey of how ideas of chance and statistical probability came to shape modern conceptions of nature, society, and the human mind.

Jaynes, E. T. (1968) "Prior Probabilities," *IEEE Transactions on Systems Science and Cybernetics* SSC-4: 227–241. Reprinted in R. D. Rosenkrantz (ed.), *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*, Dordrecht: Reidel, 1983, pp. 114–130. A concise account of Jaynes's views on Objective Bayesianism and the principle of maximum entropy.

Jaynes, E. T. (2003) *Probability Theory: The Logic of Science*, Cambridge: Cambridge University Press. An encyclopedic summation of Jaynes's approach to Objective Bayesianism, data analysis, and the role of the principle of maximum entropy.

Jeffrey, R. C, (2004) *Subjective Probability: The Real Thing*, Cambridge: Cambridge University Press. A short, engaged and engaging elaboration and defense of the subjectivist interpretation of probability.

There are any number of textbooks, of greater or lesser mathematical sophistication, on probability theory, statistical inference, and information theory.

## References

Bar-Hillel, Y. and R. Carnap (1953) "Semantic Information," *British Journal for the Philosophy of Science* 4: 147–157.

Cox, R. T. (1946) "Probability, Frequency and Reasonable Expectation," *American Journal of Physics* 14: 1–10.

Cox, R. T. (1961) *The Algebra of Probable Inference*, Baltimore, MD: Johns Hopkins University Press.
Good, I. J. (1950) *Probability and the Weighing of Evidence*, London: Charles Griffin.
Good, I. J. (1966) "A Derivation of the Probabilistic Explication of Information," *Journal of the Royal Statistical Society, Series B (Methodological)* 28: 578–581.
Hacking, Ian (2006) *The Emergence of Probability: A Philosophical Study of Early Ideas About Probability, Induction, and Statistical Inference* (second edition), Cambridge: Cambridge University Press.
Khinchin, A. I. (1953) "The Entropy Concept in Probability Theory" [Russian], *Uspekhi Mathematicheskikh Nauk* 8(3): 3–20. Translated in Khinchin, *Mathematical Foundations of Information Theory*, New York: Dover, 1957.
Milne, P. (2012) "Probability as a Measure of Information Added," *Journal of Logic, Language and Information* 21: 163–188.
Morgan, C. and E. Mares (1995) "Conditionals, Probability, and Non-triviality," *Journal of Philosophical Logic* 24: 455–467.
Popper, K. R. (1957) "A Second Note on Confirmation," *British Journal for the Philosophy of Science* 7: 350–353.
Popper, K. R. (1959) *The Logic of Scientific Discovery*, London: Hutchinson.
Shannon, C. E. (1948) "A Mathematical Theory of Communication," *Bell System Technical Journal* 27: 379–423, 623–656. Reprinted in C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, Urbana, IL: University of Illinois Press, 1949.
Williams, P. M. (1980) "Bayesian Conditionalisation and the Principle of Minimum Information," *British Journal for the Philosophy of Science* 31: 131–144.