On thinking probabilistically

M.E. McIntyre

Centre for Atmospheric Science, Dept. Applied Math. & Theor. Phys., Univ. of Cambridge, UK.
In *Extreme Events* (Proc. 15th 'Aha Huliko'a Workshop), SOEST, U. of Hawaii, 2007, pp. 153–161.

Abstract. This is about a personal journey starting from my lifelong skepticism about statistical significance tests — perhaps the most mis-applied of all mathematical theories, especially as regards extreme events — toward a new clarity and power in the use of probability theory and a clear resolution of old dilemmas about subjectivity versus objectivity. There is little original thought here. Rather, the idea is to pull together some rudimentary threads, often seen as unrelated, from mathematics, biology, experimental psychology, and information theory.

Introduction

My title is a joke, though a serious one. If one thinks carefully about what is involved, in the spirit of our "assembly that seeks into the depth of a matter", then one sees that in a certain sense the title is pleonastic, like "singing vocally". The reason is that, at the most fundamental levels — and I mean fundamental biologically as well as mathematically — there is no such thing as deterministic thinking. Our very thought processes, including mathematical thought processes, are fundamentally and inherently probabilistic.

I'll argue that this point, appropriately developed, throws light on the difficulties and controversies among statisticians and other scientists, whether about commonplace events or extreme events. They include the old "Bayesian versus frequentist" controversy between Harold Jeffreys and Ronald Aylmer Fisher and the "frequentist versus personalist" dichotomy stressed in Jay Kadane's interesting talk. They include related issues of subjectivity versus objectivity.

On primordial biological levels the point is elementary as well as fundamental. It's an almost trivial aspect of what has long been known about how biological systems work, including our own brains. The ubiquitous protein molecules called allosteric enzymes are logic elements (e.g., *Monod*, 1971). But they interact in massively-parallel information-processing "circuits" whose very "wiring" is probabilistic, indeed stochastic. Brownian motion — thermal fluctuation on picosecond timescales — connects those logic elements together in a fundamentally noisy way. That of course is why, given the mechanical strengths of chemical bonds including hydrogen bonds, life can exist only in a rather narrow temperature range.

So the textbook view of brain function, in which neurons with their tree-like dendrites and spines are viewed as deterministic adding machines, taking weighted sums of synaptic inputs (e.g., *Warwick*, 1997), is naïve and in some ways may even be profoundly wrong. Neurophysiological research in recent decades has shown that a single neuron, far from being a simple adding machine, is a highly subtle and complex information-processing system (e.g., *Crick*, 1994; *Koch*, 1999), a massively-parallel stochastic computer in its own right. The naïve neurons-and-synapses picture, more characteristic of artificial neural networks than of real ones, is showing us, at most, the surface of a vast stochastic-computational ocean about which little is understood in detail.[1]

But what about the abstract mathematical level? We have what looks at first sight like a perplexing paradox, a complete intellectual impasse. It seems to have stumped even so great a thinker as Roger Penrose, who argues — fascinatingly but wrongly, in my humble opinion[2] — that the only way out of the impasse is to suppose that the brain performs an unknown kind of quantum-gravity computation (*Penrose*, 1989, 1994).

What then is this impasse? The problem that exercised Penrose, which I too find fascinating, is the problem of how our exquisite sense of mathematical precision, of "unassailable mathematical truth", the Platonic beauty and precision of simple mathematical curves and other deterministic constructs, can possibly emerge from the actions of our tens of billions of interconnected neurons, subsisting in their metaphorical ocean of ther-

[1] One might regard the phrase "biological determinism" as another joke — this time an incongruous juxtaposition — were it not for its incessant repetition, mostly by non-biologists, and the political exploitation thereof (e.g., *Segerstråle*, 2000).

[2] Appendix to *McIntyre* (1997b); see also the sections discussing "acausality illusions" and compare them with p. 386 of *Penrose* (1994). There I *can* lay some claim to original thinking, because the insights into brain function arose from considering how music works, and its deepest connections with mathematics, in a way that I've never seen discussed elsewhere.

mal fluctuations. How, indeed, can there emerge so precise, austere, and deterministic a way of thinking as Aristotelian logic and its further developments — the unequivocal logical thinking on which we all rely, and use to build and verify our own perfectly precise and deterministic mathematical knowledge, such as knowing that there's an infinity of prime numbers?

I'll argue that both experimental psychology and evolutionary biology have something profound to say about this. But so too, it turns out, does mathematics itself, in a startling piece of intellectual bootstrapping, as summarized in the next two sections. It does so in a way that beautifully meshes with the insights from experimental psychology and evolutionary biology.

Remarks on the foundations of mathematics and probability theory

Mathematics tells us that our thought processes are "fundamentally and inherently probabilistic" in a different and entirely abstract sense that seems at first sight far removed from, and independent of, the molecular-level biological details. There is a sense in which the abstract axioms of probability theory are built into our brains. Understanding this turns out not just to be interesting in its own right, but also to be a major step toward resolving the old controversies about subjectivity versus objectivity in statistical inference.

What leads me to say such a thing? The first hint, and thrill of surprise, came when I learned that one of our most eminent mathematicians, the Fields medalist David Mumford, has gone so far as to propose that the very foundations of mathematics should be reformulated on a stochastic basis (*Mumford*, 2000). Mumford in turn cites as inspirational a book I'd never heard of, unpublished at the time and now published only posthumously, by Edwin T. Jaynes (*Jaynes*, 2003). I notice that *Kadane* (2007) lists it as "an idiosyncratic book by a controversial figure".

On turning to Jaynes' book I found, to be sure, some polemics but more importantly a clear, simple and far-reaching conceptual framework in which not only can Aristotelian logic be seen as part of probability theory but in which, under surprisingly weak qualitative assumptions, all the rules of probability theory itself can be deduced, cleanly and uniquely, from a single primordial idea that incorporates in a natural way the fact that *information* is involved. This is done using the theorems of Richard T. Cox (*Cox*, 1946), making me wonder why I was not taught those theorems as an undergraduate, instead of coin-tossing and such.

The idea is simply that, for given background knowledge or information Z, our brains must be able to attribute a degree of plausibility to any new statement, proposition or hypothesis A with which they are confronted. From this single primordial idea emerges the whole edifice of probability theory — along with an enhanced understanding of how to use it — provided only that we assume that "degree of plausibility" is somehow measured by a smoothly-variable real quantity

$$P(A|Z) \tag{1}$$

that's well defined and behaves *qualitatively* in a way that's both self-consistent, and consistent with the most rudimentary common sense (next section).

So although we begin, at this primordial level, by reading the symbol $P(A|Z)$ as the subjective "plausibility" that A is true given that Z is true, we find under weak qualitative assumptions that such symbols are mathematically indistinguishable from probabilities. That is, to the extent that our brains can assess plausibilities in some such way, they must be probabilistic devices in a high-level abstract sense — only remotely and indirectly related to the molecular-level sense noted previously — of being compelled to work with quantities P that turn out to be nothing but probabilities in the standard *quantitative* mathematical sense.

I say "primordial" and "compelled" advisedly. It hardly needs saying that the ability to make plausibility assessments of the kind in question, consciously or unconsciously, are matters of life and death and must be evolutionarily ancient.

For survival's sake, the brain must assess the plausibility that something is edible, or hostile, or whatever; and it is a biological advantage to make such assessments in a self-consistent way. As far as evolution and natural selection are concerned, departures from self-consistent calculation are mistakes. They reduce the chances of survival. So approximate self-consistency, alongside computational speed, will have been strongly selected for.[3] None of this, incidentally, has to do with the separate issues of linguistic capability and consciousness. As the anthropologist–philosopher Gregory Bateson once wrote (*Bateson*, 1972),

> "No organism can afford to be conscious of matters with which it could deal at unconscious levels."

[3]But not *exact* "algorithmic soundness", in Penrose's sense. The reader who finds it startling that there's any abstract mathematical property that can be selected for, in the biological or Darwinian sense, may find some interest in other examples. One is that, for clear reasons of survival, *prime numbers* have evidently been selected for in the case of, for instance, the *Magicicada* genus, the 13- and 17-year "periodical cicadas" of eastern North America (http://en.wikipedia.org/wiki/Magicicada#Life_cycle & refs.). One might say metaphorically that mother Nature is a mathematician and that this is no surprise, because mathematics is just a way of saying what is self-consistent.

That of course applies to ourselves just as much as to other living organisms, as I have discussed in detail elsewhere (*McIntyre*, 2000).

Incidentally, anyone who doubts that we have unconscious mathematics, and that we can do rather precise unconscious calculations, need only spend half a second looking at the "walking lights" display on my home page `//www.atm.damtp.cam.ac.uk/people/mem/`. This classic of experimental psychology (e.g., *Johansson*, 1975) clearly shows, among other things, the brain's unconscious mastery of Euclidean geometry. Twelve bright dots or splotches move over the retinas of your eyes, forming a pattern in 3-dimensional spacetime, 2 spatial dimensions and 1 time dimension. The brain fits to these sparse data a model of a certain piecewise-rigid motion in 4-dimensional spacetime, representing a person walking in the dark with light sources at his or her principal joints. For anyone with normal vision this perceptual phenomenon is highly robust and highly repeatable. Of course the brain must also be using something like Bayesian inference with unconscious priors, as is often pointed out these days (e.g., *McIntyre*, 1997a; *Jaynes*, 2003, §5.4) — but now I'm getting ahead of myself.

The consistency requirements

What exactly are those surprisingly weak qualitative assumptions about $P(A|Z)$ regarding well-definedness, self-consistency, and common sense? The most careful discussion I have seen is that of *Van Horn* (2003); see also *Cox* (1946) and chapters 1 and 2 of *Jaynes* (2003). I'll give a brief sketch to show the essence.

Well-definedness and self-consistency imply, for instance, that the same value of $P(A|Z)$ must be obtained regardless of the way it's calculated from the statements A and Z. Both statements can be arbitrarily complicated Boolean expressions. They could be sets of simpler statements connected by "and" operators, or they could be any Boolean expressions at all, provided only that whatever appears to the right of the vertical bar is not self-contradictory, i.e., not tautologically false. That would render the symbol $P(A|Z)$ meaningless, as a measure of the plausibility of A given that Z is true.

The value of $P(A|Z)$ must be independent of any Boolean rearrangements of A and Z and of the way in which their information content is packaged and labelled, as long as the information content remains the same. The fact that there are many different Boolean expressions with the same information content provides one set of constraints on the functional form of $P(A|Z)$.

Of course some packagings of information may be more convenient than others. For instance it may be convenient, even though not essential, to make Z denote the currently available background information, as already hinted. So, for now, let us think of Z as containing all the statements already known to be true at some time, while A could be a new set of statements that might or might not be true and whose plausibility we are therefore interested in assessing.

Thus we might want Z to be a large set of true statements along the lines of "I am a member of a particular species inhabiting a certain jungle" together with statements expressing everything I've learned from my experiences to date, it being irrelevant how that information is represented, e.g., verbally or non-verbally, or consciously or unconsciously. And at the time considered, statement A might be, for instance, the statement (represented somehow) that "the thing moving in front of me is a potential mate".

With A more or less plausible, though uncertain, I might need to assess the plausibility of other statements as well, such as a statement B that "the fastest escape route is such-and-such". And I might need to calculate, consciously or unconsciously, the value of $P(AB|Z)$ where AB is shorthand for the Boolean expression "A and B", so that $P(AB|Z)$ means the plausibility, given Z, that A and B are true simultaneously.

Relevant to this last is the value of $P(B|AZ)$, which can differ from $P(B|Z)$. For if A turns out to be true then the implications about possible escape routes might change. Once $P(A|Z)$ has been found, $P(B|AZ)$ is relevant and $P(B|Z)$ irrelevant to any calculation of $P(AB|Z)$. Therefore we assume

$$P(AB|Z) \;=\; \text{function of } \; P(B|AZ),\, P(A|Z) \qquad (2)$$

which in turn, however, implies because of Boolean commutativity $AB = BA$ that

$$P(AB|Z) \;=\; \text{function of } \; P(A|BZ),\, P(B|Z) \qquad (3)$$

with the same function as in (2) — both (2) and (3) holding for all Boolean expressions A, B, Z, AZ, and BZ as long as the last three are not self-contradictory.

In order to make sense as measures of plausibility, P values must vary continuously and monotonically with the degrees of plausibility. It is convenient to take them as monotonically increasing. Consistency then demands that the function on the right-hand sides of (2) and (3) is continuous and monotonically increasing in both its arguments. Suppose for instance that the information Z is updated to Z' such that the second argument on the right of (2) increases while the first does not. That is, the update makes A more plausible but is irrelevant to the plausibility of B given A, or more precisely

$$P(A|Z') > P(A|Z) \;\text{ while }\; P(B|AZ') = P(B|AZ) \quad (4)$$

which, we're assuming, has to imply

$$P(AB|Z') \geqslant P(AB|Z) \tag{5}$$

with equality only when B and therefore AB represent impossibility. Similar assumptions are made when any of the other arguments on the right of (2) and (3) are singled out as increasing; and continuity demands that small increases in those arguments produce small increases in $P(AB|Z)$.

If we now use Boolean associativity $A(BC) = (AB)C$ we can *prove* from the above that the function on the right of (2) and (3) can be taken without loss of generality to be a simple arithmetical product, so that

$$P(AB|Z) = P(B|AZ)P(A|Z) = P(A|BZ)P(B|Z) \tag{6}$$

which is the standard, quantitative "product rule" of probability theory. Notice that the second equality expresses what is usually called Bayes' theorem or Bayes' rule, with suitable choices of A, B, and Z. If we demand that (6) be compatible with the limiting cases of certainty and impossibility, we see at once that P values must always run between 0 and 1,

$$0 \leqslant P \leqslant 1 \tag{7}$$

with 0 representing impossibility and 1 certainty. And if finally we assume that

$$P(\bar{A}|Z) = \text{function of } P(A|Z) \text{ alone} \tag{8}$$

as the only reasonable relation between the plausibility of A and that of its Boolean negation $\bar{A}$, i.e., "not A", with the function monotonically decreasing, then we can *prove* with no further assumptions that the standard sum rule

$$P(A|Z) + P(\bar{A}|Z) = 1 \tag{9}$$

holds as well, for general A and Z. We now have the complete set of rules defining probability theory. The rest of probability theory follows from (6) and (9).

The proofs are far from trivial, though not difficult if one further assumes that the functions in (2), (3), and (8) are differentiable. Proofs can be given under weaker assumptions (e.g., *Van Horn* 2003 & refs).[4]

From (6) and (9) we may deduce, after a few lines of manipulation following the rules of Boolean algebra, the "extended sum rule"

$$P((A \cup B)|Z) = P(A|Z) + P(B|Z) - P(AB|Z) \tag{10}$$

where $A \cup B$ means "A or B" and satisfies $\overline{(A \cup B)} = \bar{A}\bar{B}$. This result is often visualized by set-theoretic Venn diagrams, reminding us of the Kolmogorov approach to probability theory and helping to check the correctness of the Boolean manipulations.

Let us summarize. Remarkable though it is, we have seen that starting from the seemingly vague and subjective notion of "plausibility" — and, given some reasonable universe of discourse,[4] relying solely on (2)–(5)ff., (8), differentiability, and the Boolean algebra of statements representing *information* of any kind — one has no choice but to arrive at the quantitative rules (6) and (9) of probability theory. Some authors including Van Horn use "Cox's theorem", singular, to indicate (6) and its proof together with (9) and its proof.

I should say that "no choice", though substantially correct, is not quite literally correct. In the foregoing sketch I have glossed over the technicality that there is, actually, some choice, though only in a trivial sense. For instance one may rescale everything such that P values run between 0 and $P_{\max}$ where $P_{\max}$ is any positive real number. But then the right-hand sides of (7) and (9) must be replaced by $P_{\max}$, and the second and third members of (6) multiplied by $P_{\max}^{-1}$. This is a mere "coordinate change" that makes no difference to the content of the theory.

Another such change, more general but similarly unimportant, is to replace P by a continuous monotonic function $Q = f(P)$. Again this just complicates the superficial appearance of the rules without changing their content. The rules are the same apart from writing $f^{-1}(Q)$ wherever P appears in (6)–(10). These are issues of entirely the same kind that led to the Kelvin scale as the natural temperature scale in elementary thermodynamics, and need not concern us further. If one wishes, one may think of (6) and (9) as the natural "canonical forms" of the rules defining probability theory or probability calculus, as it is also called.

Conditioning statements are primordial

It is noteworthy that conditioning statements such as Z, or whatever comes after the vertical bar, automatically appear as a natural, inevitable, and inherent part of the theory. Concealment of all the conditioning statements — as seems usual in traditional undergraduate courses on probability theory that start with coin-tossing and such — misleadingly suggests that conditioning statements are an optional add-on to be brought in later. On the contrary, it is clear from the above that they are elementary, fundamental, and primordial. They've been shown objectively, by the above arguments, to have a key status in the conceptual framework.

[4] With these and with (5) itself there are technical issues such as "universality" and "refinability" that amount to assuming a sufficiently large universe of discourse, or event space, as noted by Van Horn and others. Thus with dice, for instance, one must recognize the possibility that the number of sides $\to \infty$. Plainly there must be a large enough supply of independent statements $A, B, \ldots$ in terms of which to express assumptions like (5). We may also note that some philosophers reject Boolean algebra. Of greater practical importance, in science at least, is care over the limiting processes that lead to continuous probability distribution functions and making, for instance, appropriate use of group theory (e.g., Jaynes' Chapter 12) to ensure "coordinate independence" in the manner long familiar in physics and chemistry.

That key status is underlined by the restriction that conditioning statements must not be self-contradictory. In other words, conditioning statements must have *some* information content. That makes sense, because no kind of plausibility or probability can be meaningful in an information vacuum. In the traditional coin-tossing problem there is, of course, some rather complex background information, all concealed in a conceptual dustbin labelled "fair coin".

Making conditioning statements explicit blows away all kinds of difficulties. Take for instance a famous problem that, despite its extreme simplicity, often fools even scientifically trained people (e.g., *Krauss*, 2001; further references and comments in endnote 35 of *McIntyre*, 1997a, for instance about the classic work of Amos Tversky and Daniel Kahneman on cognitive illusions). Handicapped by my coin-tossing training, I had to think hard the first time I got it clear. This is the "three cards" or "Monty Hall" problem, of which Carl Wunsch reminded us at the Workshop (*Wunsch*, 2007) and which will come up again in my concluding remarks.

A card game is played by two people whom I'll call Monty and Mike. It is played by the following rules. Monty and Mike trust each other to follow the rules.

Monty puts an ace and two ordinary cards face-down in a row, noting the position of the ace but keeping it hidden from Mike, who has to guess where the ace is by fingering the back of one card. Monty then has to remove an ordinary card from another position and show it to Mike, who then has to bet on whether to persist with the original guess or to switch to the other face-down card — that is, to bet on which of the two remaining face-down cards is more likely to be the ace. Psychologists have found that in the role of Mike most people intuitively feel that the two cases are equally probable, and that there's no point in switching. However, on reasonable assumptions it's twice as probable that the other face-down card is the ace. That is, it would be far better to switch.

Why did I, for one, have to think hard to get this simple point clear? As we'll see in a moment, the point becomes clear as soon as one equips oneself with (6), (9), and their conditioning statements. Cox's theorems tell us that these two rules and the rules of Boolean algebra must be enough for the purpose. So if I'm Mike, all I need is to be clear what background information Z is in my possession about Monty and the card game. It is this background information Z that's concealed by what I'm calling the traditional coin-tossing training, the traditional "frequentist" training in probability theory. I remember lecture after lecture with never a conditioning statement in sight. Probability was presented as an absolute: *the* probability of this or that. This left me with no safe way to think about the three cards problem beyond a laborious enumeration of all the possibilities, including the fingering of all three positions, with careful checking that no possibility had been overlooked. Of course there's nothing wrong with frequentist thought-experiments, in their place; indeed, computer-aided Monte Carlo techniques provide very useful tools in some problem areas. Nevertheless, leaving Z implicit felt to me like groping in the dark.

Making Z explicit was like turning on the lights. Let's label the positions of the three face-down cards successively as 1, 2, and 3. Since Z represents *my* background information, as distinct from Monty's, it tells me nothing about where the ace might be. So if A_n is the statement that the ace is in position n ($n = 1, 2, 3$), it's reasonable to take $P(A_n|Z) = 1/3$ for each n. So by symmetry I may as well go ahead and finger the position $n = 1$, from here on treating that fact as updating my background information to, say, Z'. Since fingering a card changes nothing else, I have $P(A_n|Z') = 1/3$ for each n. My background information does, however, include the rules of the game. So it tells me that Monty knows where the ace is and will never remove it. So, denoting by R_n the statement that Monty removes an ordinary card from position n, I can reasonably take

$$P(R_2|A_1Z') \;=\; P(R_3|A_1Z') \;=\; 1/2 \qquad (11)$$

and

$$P(R_2|A_3Z') \;=\; P(R_3|A_2Z') \;=\; 1 \qquad (12)$$

whereupon it becomes obvious by symmetry — and in any case verifiable from (6), (9) and (10) — that after Monty removes an ordinary card (i.e., after R_2 or R_3 eventuates) the probability that I fingered the ace is still 1/3 and therefore that I didn't finger it 2/3, by (9). More precisely, $P(A_1|R_2Z') = P(A_1|R_3Z') = 1/3$ and therefore, by (9), $P(A_3|R_2Z') = P(A_2|R_3Z') = 2/3$.

To the extent that we regard P values like those in (11) and (12), and the values $P(A_n|Z) = P(A_n|Z') = 1/3$, as properties of the background information we may usefully call them *prior probabilities* or *priors* — even though we haven't explicitly used Bayes' rule. Of course all those P values are assumed values. They all involve subjective judgement, however reasonable that judgement may seem.

But is it always reasonable? Suppose I suddenly recall that Monty has some bias such that (11) becomes

$$P(R_2|A_1Z') \;=\; q\,, \quad P(R_3|A_1Z') \;=\; 1-q \qquad (13)$$

where $q \neq 1/2$. For instance I might recall that Monty has a form of Tourette's syndrome that compels him to remove the card as near as possible to the card I fingered, making $q = 1$. Then if R_3 eventuates, I can be certain of A_2. That is, $P(A_2|R_3Z') = 1$. For general q it can be verified from (6), (9) and (10) with, as always, attention to the rules of Boolean algebra,

first that $P(R_3|Z') = P(R_3|A_1Z')\ P(A_1|Z') + P(R_3|A_2Z')\ P(A_2|Z') = (2-q)/3$ and then, after further use of (6), that $P(A_1|R_3Z') = (1-q)/(2-q)$ and $P(A_2|R_3Z') = 1/(2-q)$. Setting $q = 1/2$ recovers the respective values 1/3 and 2/3 appropriate to (11).

Bayesian inference and Platonic forms

I want to return to our *unconscious* mathematics, as manifested by the walking lights demonstration, before coming to some final remarks about extreme events, the subjectivity-versus-objectivity issue, and the old disputes between frequentists, Bayesians and others.

How confident can we be that natural selection will produce something operationally equivalent to (6) and (9), to good approximation? Although we are far from being able to verify this directly from neurophysiology, we may note all the other examples of how natural selection tends to optimize functionality. The cicada example of footnote 3 is only one among countless others (the point, there, being that prime numbers of seasonal cycles tend to minimize encounters with predators). Anyone who has ever taken the slightest notice of biological phenomena must surely be impressed by the approach to optimal solutions that we see wherever we look in the living world. And Cox's theorems tell us that (6) and (9) are themselves functionally optimal in a very strong sense.

The streamlined shapes of high-speed fish and birds provide further examples of near-optimization by natural selection. The beautiful curves describing the shapes of efficient airfoils reflect an aerodynamic functionality that's close to optimal. Precisely optimal shapes are not of course attained by real fish and birds, but appear to be well approximated in many cases (e.g., *Lighthill*, 1975, p. 32). Such shapes, with their simplicity and infinite smoothness, are examples of Platonic geometric forms. The world of Platonic forms, the idealized world of mathematical wonders where perfect circles are truly perfect, and curves can have infinite numbers of derivatives, reminds us in turn of Penrose's impasse.

But once again I'm getting ahead of myself. What have Platonic geometric forms got to do with the rules of probability theory? There is an interesting answer. Both have a great deal to do with the primordial kind of statistical inference we call visual perception.

We have already met this point in the walking lights demonstration. Perception works by model-fitting using unconscious priors, some of them coming from genetic memory. Insofar as genetic memory has also equipped us with (6) and (9) it seems reasonable to suppose, indeed almost inevitable, because of natural selection, that the visual brain's unconscious model-fitting process must, as already hinted, be operationally equivalent to Bayesian inference.

Cox's theorems tell us that anything else would imply inconsistency in using the available visual information, which would have been selected *against*.

Thus, in the case of the walking lights, the model-fitting process must to good approximation be equivalent to taking, say, A in (6) to represent the data and B to represent a hypothesis, or candidate model, while Z contains any relevant background information. So Z would consist of a large number of statements that might include, for instance, "my eyes are open and I'm looking at something near the top of Michael McIntyre's home page". A would be a statement asserting that twelve bright dots are moving across my retinas, tracing certain smooth curves, or mostly-smooth curves, in 3-dimensional spacetime. B would be a statement defining one of a combinatorially large set of hypotheses, or candidate models, to be fitted to the data.

For instance if we interdistinguish the B statements by subscripts, $B = B_i$ say, varying discretely or continuously, then there might be a large subset $\mathbf{I}$ of i values such that the corresponding B_i all state that the twelve bright dots are such as could originate from glow-worms, or other light sources, moving in various ways on a wall, or otherwise defining smooth or mostly-smooth curves in 3-dimensional spacetime. Let's call the corresponding candidate models "planar models". For another subset $\mathbf{J}$, the corresponding B_j with $j \in \mathbf{J}$ might all state that the moving dots are such as could originate as a perspective view of light sources at the pivots of a jointed skeletal structure, in piecewise-rigid reciprocating motion tracing smooth or mostly-smooth curves in 4-dimensional spacetime. Let's call the corresponding models "non-planar models".

The candidate models are indeed *models*, i.e., are hypothesized to be partial and approximate representations of reality. Therefore, despite our Platonic sensibilities, the spacetime curves so defined are best thought of as being slightly fuzzy. So we may as well include the fuzziness in each model specification. Indeed, the repertoire of candidate models should include models with different amounts of fuzziness. For the particular model specified by statement B_j we can read $P(A|B_jZ)$ as measuring the goodness of fit of that model to the data specified by A, maximizing when the j value gives the best fit possible. Rewriting the second part of (6) in the usual Bayesian manner for this purpose, we have

$$P(B_j|AZ) = \frac{P(A|B_jZ)\,P(B_j|Z)}{P(A|Z)} \qquad (14)$$

for $j \in \mathbf{J}$, and similar equations for B_i with $i \in \mathbf{I}$. Since, in the usual Bayesian manner (*Kadane*, 2007), we are focusing on just the one dataset — i.e., are regarding the statement A as fixed — we are free to ignore the

denominator when comparing models and to pay attention only to relative P values in the numerator.[5]

In the present case, it's obvious that either kind of model, planar or non-planar, can fit the data equally well. So the outcome depends entirely on the relative values of $P(B_i|Z)$ or $P(B_j|Z)$, the brain's *unconscious priors* for the various models. The perceptual phenomenon actually experienced, by everyone with normal vision, shows that the non-planar models are overwhelmingly favored, at least in cases where the 4-dimensional motion resembles any human or animal skeletal motion. That is, the unconscious priors must be such that $P(B_j|Z) \gg P(B_i|Z)$ for $i \in \mathbf{I}$, $j \in \mathbf{J}$.

How the brain actually uses its multi-level, massively parallel computational abilities, from allosteric enzymes upward, to carry out such computations, and exactly how it narrows down a combinatorially large repertoire of candidate models — or indeed just what that repertoire consists of — are all largely unknown. In view of the combinatorial largeness it is surely safe to assume, however, that Occam's-razor principles are involved. As I discuss elsewhere (*McIntyre*, 1997a) we glimpse some of these from phenomena such as perceptual grouping, and others through our visions of Platonic forms such as smooth spatial and spatiotemporal curves.

Figure 1, another classic of experimental psychology, directly demonstrates the unconscious priors that express the Platonic by favoring simple, smooth curves. Anyone with normal vision who stares at this figure sees a beautifully smooth curve, ghostly yet sharp and precise, fitted to the inner ends of the black segments. Experimental psychologists call these curves "illusory contours". The figure is from *McIntyre* (1997b); other examples can be found in practically any book on vision, e.g., *Crick* (1994, p. 47).

The brain chooses a simple, smooth spatial curve. So it must be using something operationally equivalent to Bayesian computations to solve, unconsciously, an extremum problem in the calculus of variations. This happens for anyone with normal vision even though other, less smooth curves would fit the inner ends of the segments equally well. It is evident that an Occam's-razor principle must be involved. An unconscious choice of priors favors the simplest possible object outlines consistent with retinal data.

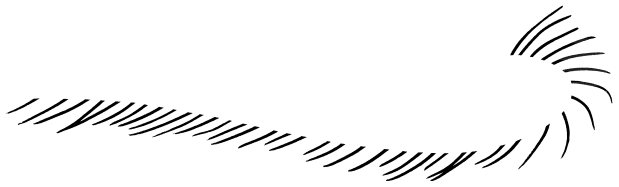

Figure 1. Demonstration of an illusory contour grazing the inner ends of the black segments (see text). In constructing the contour, which does not exist physically on the paper or screen, the visual system is unconsciously solving a Bayesian problem that is also an extremum problem in the calculus of variations, minimizing some norm involving rates of turning of the tangent to the contour.

There are auditory counterparts in music, as I have discussed elsewhere (*McIntyre* 1997a,b). The Platonic is, indeed, awesomely, "already there" (e.g., *Penrose*, 1989, 1994) and is something that artists, mathematicians, and many others have always recognized. Being in genetic memory, it must be evolutionarily ancient. It has been "there" from time immemorial, many tens of millions of years at least.

Concluding remarks

The foregoing reminds us of what is perhaps the most deep-seated and elusive difficulty with probabilistic thinking — the extreme intimacy and complex interplay between reason and intuition, much of it beyond conscious reach. In the parlance increasingly popular in the AI and neuroscience communities today, the brain is a massively-parallel Bayesian machine (e.g., *Ghahramani*, 2004) with a vast and ever-changing, context-sensitive web of unconscious assumptions as priors. So when we use our brains to think consciously about probabilities, we have to contend with powerful intuitions from the unconscious probabilistic thinking already present — if "thinking" is the right word. In particular, we can't know about all the unconscious priors. It's no wonder that the whole subject area has been a quagmire of endless debates over what's subjective and what's objective, over what various statistical tests "really mean", and in a wider community over even the simplest problems like that of the three cards and that of the "prosecutor's fallacy", which latter has led to terrible consequences such as unsafe murder convictions.

Cox's theorems seem to me to provide crucial guidance in negotiating that quagmire. They establish that

[5] By long-established convention, the goodness of fit $P(A|B_j Z)$ is called the "likelihood" (of the model specified by B_j as a generator of the given data specified by A). Since one is interested only in the relative goodness of fit when comparing any two models, one usually deals with "likelihood ratios" such as $P(A|B_j Z)/P(A|B_i Z)$, or their logarithms. A set of candidate models is itself, by association, called "a likelihood". The rationale is that with A fixed one is regarding a symbol like $P(A|B_j Z)$ as a function of the B_j. So if one hears a statistician ask another statistician "what's your likelihood?" it probably means "what set of models are you trying to fit to this particular dataset?" Such is human language. Model-fitting of this kind is supremely important because it represents the functioning not only of perception but also of the systematic and more conscious extension of it that we call science. Chapter 3 of *MacKay* (2003) and chapter 4 of *Jaynes* (2003) give valuable discussions and examples.

if one accepts Boolean algebra — and that plausibilities or probabilities can be measured as smoothly-varying, real-valued functions — then there is nothing subjective about the framework of probability theory expressed by (6), (9) and (14) and their mathematical consequences. So subjectivity comes *entirely* from the choice of information to be considered and from any failure to make it explicit, and to handle it in a consistent way.

In particular, there are always conditioning statements, some of which represent background information. There are always, therefore, some priors that must be estimated or somehow chosen. Even in the three-cards problem, as I hope I made clear on page 157, it is valuable just to say explicitly what the background information and priors are taken to be, the priors being the assumptions that $P(A_n|Z) = P(A_n|Z') = 1/3$ together with (12), and (11) or (13). The reasonableness of the priors can be better judged once the information is made explicit. It comes down to solving the problem with one's eyes open rather than closed.

That kind of clarity seems to me to be no luxury, and not just for murder trials. It's an urgent necessity in more complicated problems, too, such as how to think rationally about extreme events affecting thousands of people. Indeed the choice of background information to be considered, and the corresponding choice of priors, can now be seen as a necessary task in any attempt at probabilistic reasoning and inference that aspires even to self-consistency, let alone to objectivity. And the old hardcore frequentist prohibition, forbidding scientists to make priors explicit, now looks more and more like commanding us to go about our business blindfolded.

Even when the background information is in some sense minimal — in some sense close to an information vacuum — one still has the challenge of choosing so-called "ignorance priors" in as consistent and objective a way as possible. Such problems can to some extent be clarified (e.g., *Jaynes*, 2003, chapter 12) albeit far from wholly resolved (e.g., *Kass and Wasserman*, 1996) by considerations of coordinate independence and group invariance. Another approach is to maximize the Shannon information entropy and appears to have worked well in certain problems, such as image processing. But that, too, has turned out to be far from universally applicable (e.g., *MacKay*, 2003, Ex. 22.13) if only because of inconsistencies with (14) (*Seidenfeld*, 1987). As Van Horn puts it, ignorance for this purpose is a slippery concept and, in practice, one is never completely ignorant (*Van Horn*, 2003, end of §4), taking unconscious knowledge into account. And if one thinks one is making no assumptions — as with the hardcore frequentists' mantra "let the data speak for themselves" — then it means only that all one's assumptions are unconscious. Besides, in real problems it may be necessary, and desirable, to include overtly subjective or "hunch" elements in the priors assumed, as the best scientists have always done, consciously or unconsciously or both. The main need is to try to make the assumptions explicit; and a saving grace is an important "chain consistency property" (*Jaynes*, 2003, eq. (8.57)ff.) that shows how, and in what circumstances, successive applications of (14) enable priors to be improved by successive inputs of data.

As already said, frequentist thought-experiments can be very useful. Yet the harm done by the old hardcore frequentist or "ultra-orthodox" view and its dominance over undergraduate education, portraying probabilities as absolutes and increasing the risk of blunders like the prosecutor's fallacy, now seems almost reminiscent of the harm once done by the contemporaneous views called eugenics and behaviorism. If you think that's a bit harsh, consider that all those views seem to have involved similar attitudes claiming ownership of an absolute truth (such as "priors are never admissible", "genes are either good or bad", or "scientists may not study perceptual phenomena") and forbidding everyone to think outside the bounds thus set — manifestations, it seems to me, of the human "hypercredulity instinct", that ancient tribal-cohesion mechanism of whose manifestations in other areas we are so painfully aware today.

Physicists at the time — as distinct from some physicists today — managed to avoid such traps. They were helped by the sheer force of experimental evidence, such as blackbody radiation, atomic spectroscopy, and the photoelectric effect and were helped also, at first, by the mind-blowing experience of having to develop quantum theory. Max Born put it well when he wrote

> "I believe that ideas such as absolute certitude, absolute exactness, final truth, etc., are figments of the imagination which should not be admissible in any field of science .... This *loosening of thinking* [Born's emphasis] seems to me to be the greatest blessing which modern science has given to us. For the belief in a single truth and in being the possessor thereof is the root cause of all evil in the world" (*Born*, 1991).

He had a point.

Acknowledgments. I warmly thank Arieh Iserles, who first made me aware of David Mumford's thinking, and Jay Kadane who patiently tutored me on the technical distinction between "probability" and "likelihood" and on the meaning of the term "Bayesian inference" as it is understood today. I am grateful to James Maas, Stuart Dalziel, Nicholas Pinhey, Steve Lay, and Björn Haßler for their help with the walking-lights and the other displays at `//www.atm.damtp.cam.ac.uk/people/mem/`. Jennifer An-

derson, P. John Anderson, Patrick Bateson, Peter Fellgett, Zoubin Ghahramani, Geoffrey Grimmett, Frank Kelly, David MacKay, Tim Palmer, Cosma Shalizi, Bill Simmons, David Spiegelhalter, Philip Stark, Marilyn Strathern, Kevin Van Horn, and Carl Wunsch all kindly made helpful comments. I also thank the organizers for daring to invite me to a speak on a topic on which I had never worked, in which I had never acquired any special expertise, and to which I have yet to make any significant contribution. So although this has been only a personal journey it has brought to me, at least, the joy of discovery and illumination.

References

Bateson, G. L., Style, grace and information in primitive art. In *Steps to an Ecology of Mind: Collected Essays on Anthropology, Psychiatry, Evolution and Epistemology*, pp. 101–125. San Francisco, Chandler; Aylesbury, Intertext; Northvale, NJ, Jason Aronson, 510 pp., 1972.[6]

Bateson, P., Martin, P., *Design for a Life: How Behaviour Develops*. London, Jonathan Cape, Random House, 280 pp., 1999. Also Simon and Schuster, USA 2000.[7]

Born, G., Problems with limits, *Science and Public Affairs* (London, Roy. Soc.), **6(2)**, 17–25, 1991.[8]

Cox, R. T., Probability, frequency and reasonable expectation. *Amer. J. Phys.*, **14**, 1–13, 1946.

Crick, F., *The Astonishing Hypothesis*. London, New York, Simon and Schuster, 317 pp., 1994.

Ghahramani, Z., Bayesian machine learning: a tiny intro. At `//learning.eng.cam.ac.uk/zoubin/bayesian.html`, 2004.

Jaynes, E. T., *Probability Theory: The Logic of Science*, edited by G. Larry Bretthorst. Cambridge, University Press, 727 pp., 2003. Chaps. 1–3, as published, available at `//bayes.wustl.edu/etj/prob/book.pdf`; 1996 prepr. at `//omega.albany.edu:8008/JaynesBook.html`

Johansson, G., Visual motion perception. *Sci. Amer.,* **232**, June issue, 76–88, 1975.

Kadane, J. B., What is an extreme event? This proceedings, 2007.

Kass, R. E., Wasserman, L., The selection of prior distributions by formal rules. *J. Amer. Stat. Assoc.*, **91**, 1343–1370, 1996.

Koch, C., *Biophysics of Computation: Information Processing in Single Neurons*. Oxford Univ. Press, 562 pp., 1999.

Krauss, S., Some Issues of Teaching Statistical Thinking. PhD dissertation, Free University of Berlin, Chapter 2, pp. 41–85, 2001. At `www.diss.fu-berlin.de/2003/311/` — also `//math.ucsd.edu/~crypto/Monty/montybg.html` for the story of Marilyn vos Savant and the nearly ten thousand wrong answers.

Lighthill, M. J., *Mathematical Biofluiddynamics*. Philadelphia, Society for Industrial and Applied Mathematics, 281 pp., 1975.

MacKay, D. J. C., *Information Theory, Inference, and Learning Algorithms*. Cambridge, University Press, 628 pp., 2003. At `//www.inference.phy.cam.ac.uk/mackay/`

McIntyre, M. E., Lucidity and science I: Writing skills and the pattern perception hypothesis. *Interdisciplinary Science Reviews*, **22**, 199–216, 1997a. Supplementary material is at `//www.atm.damtp.cam.ac.uk/people/mem/` including animated graphics.

McIntyre, M. E., Lucidity and science II: From acausality illusions and free will to final theories, mathematics, and music. *Interdisc. Sci. Revs.*, **22**, 285–303, 1997b. Supplementary material including animated graphics is at `//www.atm.damtp.cam.ac.uk/people/mem/`

McIntyre, M. E., Lucidity, science, and the arts: what we can learn from the way perception works. *Bull. Faculty Human Devel.* (Kobe University), **7**(3), 1–52, 2000. Invited keynote lecture to the 4th Symposium on Human Development, *Networking of Human Intelligence: Its Possibility and Strategy* held in Kobe, Japan, on 4 Dec. 1999. Available at `//www.atm.damtp.cam.ac.uk/people/mem/`

Monod, J., *Chance and Necessity*, transl. A. Wainhouse. Glasgow, Collins, 187 pp., 1971.

Mumford, D., The Dawning of the Age of Stochasticity. In *Mathematics: Frontiers and Perspectives*, edited by V. I. Arnol'd, M. Atiyah, P. Lax and B. Mazur, Providence, RI, Amer. Math. Soc., 460 pp., 2000.

Penrose, R., *The Emperor's New Mind*. Oxford, University Press, 466 pp., 1989.

Penrose, R., *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford, University Press, 457 pp., 1994.

Segerstråle, U., *Defenders of the Truth: The Battle for Science in the Sociobiology Debate And Beyond*. Oxford, University Press, 493 pp., 2000. See also *Bateson and Martin* (2000) for a deeper scientific insight that transcends the debate.

Seidenfeld, T., Entropy and uncertainty. In *Foundations of Statistical Inference*, edited by I. B. MacNeill and G. J. Umphrey, Reidel, Dordrecht, 259–287, 1987.

Van Horn, K., Constructing a logic of plausible inference: a guide to Cox's theorem. *Intl. J. Approx. Reasoning*, **34**, 3–24, 2003.

Warwick, K., *March of the Machines: Why the New Race of Robots will Rule the World*. London, Random House (Century Books), 263 pp., 1997. The argument presupposes what I called the "naïve neurons-and-synapses picture, more characteristic of artificial neural networks than of real ones." With prodigious assurance the author goes on to predict that robots will be smarter than humans, whatever that means, within about half a century.

Wunsch, C., Extremes, patterns, and other structures in oceanographic and climate records. This proceedings, 2007.

[6] Gregory Bateson (1904–1980) began his career as an anthropologist. His writings contain much wisdom. His father was the genetics pioneer William Bateson (1861–1926), and Patrick Bateson (next ref.) is his second cousin once removed.

[7] This lucid and authoritative book, from acknowledged experts on the evidence from animal and human behavior, exposes the naïveté of simplistic biological determinism as well as the absurdity of the popular false dichotomizations such as "nature *or* nurture", comparable to the absurdity of "lock *or* key".

[8] Gustav Born is Max Born's son.

This preprint was prepared with AGU's LaTeX macros v4, with the extension package 'AGU^{++}' by P. W. Daly, version 1.6a from 1999/05/21.