

Should you believe that this coin is fair?

William Bialek

*Joseph Henry Laboratories of Physics and the Lewis–Sigler Institute for Integrative Genomics
Princeton University, Princeton, New Jersey 08544*

(Dated: August 30, 2005)

Faced with a sequence of N binary events, such as coin flips (or Ising spins), it is natural to ask whether these events reflect some underlying dynamic signals or are just random. Plausible models for the dynamics of hidden biases lead to surprisingly high probabilities of misidentifying random sequences as biased. In particular, this probability decays as $N^{-1/4}$, so that no reasonable amount of data would be sufficient to induce the concept of a fair coin with high probability. I suggest that these theoretical results may be relevant to understanding experiments on the apparent misperception of random sequences by human observers.

There is a large literature testifying to the errors that humans make in reasoning about probability [1, 2]. Perhaps most fundamental is the claim that people routinely detect order and hidden causes in genuinely random sequences [3]. These apparent limitations on human rationality have broad implications, not least for economics, and have attracted considerable attention in the popular press. In contrast with these results, a number of experiments indicate that humans and other animals can change their behavior in response to changes in the probabilities of stimuli and rewards, sometimes making optimal use of the available data [4, 5, 6, 7]. Similarly, many perceptual discriminations approach the limits to reliability set by noise near the sensory input [8, 9, 10], and related ideas of statistical optimization have emerged in recent work on motor control [11, 12, 13]. There is even the suggestion that if the detection of order vs. randomness is cast in the standard two–alternative format for perceptual discrimination experiments, then people can learn to perform with close to the statistically maximum reliability [14]. While nearly perfect neural processing of statistical data under some conditions could coexist with qualitative failures at similar problems under different conditions, it would be attractive to have a more unified view of the brain as an engine for probabilistic inference.

The problem of identifying genuinely random sequences has a number of subtleties that seem not to have been emphasized in the previous literature. In particular, from a Bayesian point of view our confidence that a given sequence really was generated at random depends entirely on the universe of alternative models that we are willing to consider. Here I consider a family of models that involve time–dependent biases of unknown magnitude, similar in spirit to the models that would allow optimal inference under the conditions of the experiments in Refs [6, 7], and show that an observer who tries to understand the world using these models will exhibit a surprisingly large probability of misidentifying random sequences as having small but nonzero biases. Most importantly this probability declines only as a fractional power of the sequence length, so that (for example) no reasonable human experience with the flipping of a coin

would provide sufficient evidence to induce the concept of fairness with high probability.

To be concrete let us consider data in the form of binary sequences—coin flips (heads/tails), for example, or a sequence of rewards/nonrewards from a particular class of actions. Let the observations come in a sequence labeled by $n = 1, 2, \dots, N$, and let the binary variable on each observation n be denoted by σ_n . If the process is described by a fair coin then all sequences occur with equal probability,

$$P(\{\sigma_n\}|\text{fair coin}) = \frac{1}{2^N}. \quad (1)$$

The problem faced by the observer, however, is to estimate the probability that this particular sequence was generated by a fair coin, that is $P(\text{fair coin}|\{\sigma_n\})$. Bayes' rule tells us that we can write this probability as

$$P(\text{fair coin}|\{\sigma_n\}) = \frac{P(\{\sigma_n\}|\text{fair coin})P(\text{fair coin})}{P(\{\sigma_n\})}, \quad (2)$$

where $P(\text{fair coin})$ measures the a priori probability of a fair coin and $P(\{\sigma_n\})$ measures the probability that this particular sequence will arise, averaged over all possible models that might describe it; schematically we can write

$$P(\{\sigma_n\}) = \sum_{\text{all models}} P(\{\sigma_n\}|\text{model})P(\text{model}), \quad (3)$$

and of course one of the possible models is the fair coin. Because the fair coin generates all sequences with equal probability [Eq (1)], the probability that a fair coin generated a given sequence depends on the details of the sequence only through the term $P(\{\sigma_n\})$, *i.e.* only through the average over all possible models of the sequence. Thus our confidence that our experience is described by a fair coin depends entirely on the set of alternative models that we think is appropriate to the situation.

A plausible set of alternatives to the fair coin is that each σ_n responds independently to a bias, but this bias may change over time. To specify this class of models completely requires at least three steps. First we need to describe the bias on each trial. To treat heads and tails

symmetrically we can represent heads/tails as an Ising spin $\sigma_n = \pm 1$, and measure the bias on observation n as an effective magnetic field h_n such that [15]

$$P(\sigma_n|h_n) = \frac{\exp(-h_n\sigma_n)}{2 \cosh(h_n)}. \quad (4)$$

Assuming that the bias acts on each observation independently, we have

$$P(\{\sigma_n\}|\{h_n\}) = \prod_{n=1}^N \frac{\exp(h_n\sigma_n)}{2 \cosh(h_n)} = \frac{1}{2^N} e^{-f(\{\sigma_n; h_n\})} \quad (5)$$

$$f(\{\sigma_n; h_n\}) = \sum_{n=1}^N [\ln \cosh(h_n) + h_n\sigma_n]. \quad (6)$$

A particular set of biases $\{h_n\}$ constitutes one possible model. The second step is that to average over models we will need a hypothesis about the distribution of these biases. Consider the case where the bias is on average zero (coins or circumstances that favor heads are as likely as those that favor tails), and where the presumably small fluctuations in bias are drawn from a Gaussian with root-mean-square fluctuations h_{rms} . Dynamics of the bias vs. time are described by a correlation function $C(\tau)$, $\langle h_n h_m \rangle = h_{\text{rms}}^2 C(n-m)$. If, for example, the bias tends to stay constant for runs of 10 observations, then $C(\tau)$ should be close to 1 for $|\tau| \leq 10$ and fall to zero for $|\tau| \gg 10$. It is convenient to think about a matrix \hat{C} with elements defined by $(\hat{C})_{nm} = C(n-m)$. Then the full distribution of biases has the form

$$P(\{h_n\}|h_{\text{rms}}) = \frac{1}{Z} \exp \left[-\frac{1}{2h_{\text{rms}}^2} \sum_{n,m=1}^N h_n (\hat{C}^{-1})_{nm} h_m \right],$$

$$Z = (\sqrt{2\pi}h_{\text{rms}})^N \exp \left[+\frac{1}{2} \text{Tr} \ln \hat{C} \right]. \quad (7)$$

As a third and final step we have to specify our knowledge of C and h_{rms} . To keep things simple let us assume that there is something about the situation which makes us certain about the time scales on which bias can vary, so that we know the correlation function $C(\tau)$, but we don't know for sure how strong the biases can get, so we have to average over some distribution $P(h_{\text{rms}})$.

Putting the various terms together, we can write the probability of observing a sequence $\{\sigma_n\}$ in the broad family of biased models as

$$P(\{\sigma_n\}|\text{biased}) = \int dh_{\text{rms}} d^N h_i P(\{\sigma_n\}|\{h_n\}) P(\{h_n\}|h_{\text{rms}}) P(h_{\text{rms}}), \quad (8)$$

and then

$$P(\{\sigma_n\}) = P(\{\sigma_n\}|\text{fair coin})P(\text{fair coin}) + P(\{\sigma_n\}|\text{biased})P(\text{biased}). \quad (9)$$

Thus the probability of an observed sequence arising from a fair coin is

$$P(\text{fair coin}|\{\sigma_n\}) = \frac{1}{1 + \exp(\Lambda - \mu)}, \quad (10)$$

where the log-likelihood ratio is

$$\Lambda = \ln[P(\{\sigma_n\}|\text{biased})/P(\{\sigma_n\}|\text{fair coin})], \quad (11)$$

and the threshold $\mu = \ln[P(\text{fair coin})/P(\text{biased})]$.

In the limit that h_{rms} is small we can compute the integral over $\{h_n\}$ in Eq (8) as a perturbation series. To fourth order in h_{rms} , the result is

$$\int d^N h_i P(\{\sigma_n\}|\{h_n\}) P(\{h_n\}|h_{\text{rms}}) = \frac{1}{2^N} \exp \left[-\frac{N\tau_c}{2} a h_{\text{rms}}^4 + \sqrt{N\tau_c} z h_{\text{rms}}^2 \right], \quad (12)$$

where the correlation time is defined by

$$\tau_c = \frac{1}{2N} \text{Tr} \hat{C}^2 = \frac{1}{2} \sum_n C^2(n), \quad (13)$$

$a = 1 + y\sqrt{4g/N\tau_c}$, $g = (\text{Tr} \hat{C}^4)/(\text{Tr} \hat{C}^2)$, and the variables z and y depend on the particular sequence that we observe:

$$z = \frac{1}{\sqrt{4N\tau_c}} \left[\sum_{nm} \sigma_n \hat{C}_{nm} \sigma_m - N \right] \quad (14)$$

$$y = \frac{1}{\sqrt{4gN\tau_c}} \left[\sum_{nm} \sigma_n (\hat{C}^2)_{nm} \sigma_m - \text{Tr} \hat{C}^2 \right]. \quad (15)$$

The normalization of z and y is chosen so that they each have zero mean and unit variance if the $\{\sigma_n\}$ actually are generated by a fair coin. The exponential in Eq (12) sets a scale for $h_{\text{rms}} \sim N^{-1/4}$, which becomes small at large N . Thus when we do the integral over h_{rms} in Eq (8) it will be dominated by very small values of h_{rms} , so the relevant parameter is $P(h_{\text{rms}} \rightarrow 0) = \rho$ [16]. Then

$$P(\{\sigma_n\}|\text{biased}) = \frac{\rho}{2^N} \int dh_{\text{rms}} \exp \left[-\frac{N\tau_c}{2} a h_{\text{rms}}^4 + \sqrt{N\tau_c} z h_{\text{rms}}^2 \right]. \quad (16)$$

At large N and $z < 0$ the integral is dominated by its behavior near $h_{\text{rms}} = 0$, while for $z > 0$ it is dominated by a saddle point at $h_* = (z^2/a^2 N\tau_c)^{1/4}$. When the dust settles, the large N approximation to the log-likelihood ratio becomes

$$\Lambda(z < 0) \sim \frac{1}{4} \ln \left[\frac{\pi^2 \rho^4}{4z^2 N\tau_c} \right] \quad (17)$$

$$\Lambda(z > 0) \sim \frac{z^2}{2a} + \frac{1}{4} \ln \left[\frac{\pi^2 \rho^4}{4z^2 N\tau_c} \right], \quad (18)$$

where the dependence on the observed sequence is through the value of z (and, negligibly in this limit, also through the dependence of a on y).

From Eq (10) we see that if $\Lambda > \mu$ then the probability of the sequence $\{\sigma_n\}$ being described by a fair coin is less than half. Equivalently, if $\Lambda > \mu$ then the data are more likely to be described by a biased model. If we want to make correct assignments with highest probability, then the correct decision rule is maximum likelihood [4], so that all sequences with $\Lambda > \mu$ should be labeled as biased. Obviously there is some probability that this condition is met even if the sequence in fact was generated at random. Note that for large N the variable z , which determines Λ , is the sum of many terms which are independent if the sequence really is random. Thus the distribution of z approaches a Gaussian (with zero mean and unit variance, by construction) and we can estimate the probability that $\Lambda > \mu$, which is the probability that a random sequence will be identified as biased [17].

The condition $\Lambda > \mu$ corresponds either to $z_- < z < 0$ or $z > z_+$, where

$$z_- = -\frac{\pi}{2} \left(\rho \frac{P(\text{fair coin})}{P(\text{biased})} \right)^2 \frac{1}{\sqrt{N\tau_c}} \quad (19)$$

$$z_+ = \ln^{1/2} \left[\frac{2z_+ \sqrt{N\tau_c}}{\pi \rho^2} \left(\frac{P(\text{fair coin})}{P(\text{biased})} \right)^2 \right] \quad (20)$$

Then a sequence generated by a fair coin will be assigned as biased with a probability given by

$$P_{\text{error}} \approx \int_{-z_-}^0 \frac{dz}{\sqrt{2\pi}} e^{-z^2/2} + \int_{z_+}^{\infty} \frac{dz}{\sqrt{2\pi}} e^{-z^2/2}. \quad (21)$$

For large N , $|z_-| \ll 1$ and $z_+ \gg 1$, which lets us approximate the integrals to find

$$P_{\text{error}} \sim r^2 \sqrt{\frac{\pi}{8N\tau_c}} + \frac{r}{2} \cdot \frac{1}{(N\tau_c)^{1/4}} \cdot \frac{1}{\ln^{5/8} [2\sqrt{N\tau_c}/\pi r^2]}, \quad (22)$$

where $r = \rho P(\text{biased})/P(\text{fair coin})$ and I neglect terms of the form $\ln \ln N$. At sufficiently large N the second term dominates, and we have simply

$$P_{\text{error}} \sim \left(\frac{N_c}{N} \right)^{1/4} \cdot \frac{1}{\ln^{5/8} \left[\sqrt{N}/(4\pi^2 N_c) \right]}, \quad (23)$$

where the scale is set by $N_c = r^4/(16\tau_c)$.

The power-law decay of the error probability is surprising if we have in mind the conventional problem of statistical hypothesis testing. With two specific alternative hypotheses, e.g. the fair coin and a coin with fixed (and known) bias, the log-likelihood ratio on average grows linearly with the number of examples that we observe, this growth rate being the Kullback-Leibler divergence between the two hypothesized distributions, and the resulting error probability falls exponentially. By allowing for the possibility that biases change with time over

the course of our observations—which seems plausible in many natural contexts—we construct a family of models with a number of parameters that grows as the size of the data set increases. In addition, the particular model of bias considered here allows larger data sets to be used to test for weaker biases (smaller h_{rms}), which again is plausible but very different from the simpler problem of discriminating between two fixed distributions. To summarize, if we test data for a fixed bias of known size, then the probability of mistakenly finding order in a random sequence decays exponentially. If on the other hand we test for time dependent biases of unknown magnitude, the error probability falls as a power-law. The case considered here generates a 1/4 power, but one can construct models that generate other powers as well [16].

In the case of two specific hypotheses, the exponential dependence of error probability also means that the trading between number of examples and prior probability is only logarithmic. Suppose that when fair and biased coins are equally likely a priori ($\mu = 0$), it takes N_0 examples to reach some criterion level of error. If we now imagine that truly fair coins are rare, in the conventional hypothesis testing view it will take $\Delta N_0 \propto \ln[P(\text{biased})/P(\text{fair coin})]$ additional examples to reach the same level of error. In contrast, when we search for time dependent biases with unknown magnitude, we are much more sensitive to prior probabilities, since Eq (23) predicts $N_0 \propto N_c$, or $N_0 \rightarrow N_0 [P(\text{biased})/P(\text{fair coin})]^4$. Concretely, if we can expect that genuinely random sequences constitute only $\sim 10\%$ of the events that we will see, that the typical scale of biases is $h_{\text{rms}} \sim 1$ so that $\rho \sim 1$, and that correlation times are $\tau_c \sim 10$ events, then $N_c \sim 40$ and hence even at $N = 10^4$ the error probability is $P_{\text{error}} \sim 0.25$ [18].

The sequences which are identified as biased are primarily those with large positive values of z from Eq (14). If the expected correlation function of the bias is everywhere positive (so that we don't expect oscillating biases), then this singles out sequences that have an excess of runs with multiple heads or tails in a row. Put another way, sequences which are declared to be random have fewer runs than expected in genuinely random distributions. This is in agreement with experiments showing that when human subjects are asked to generate random sequences or to assess sequences for randomness, they behave according to a “representativeness heuristic” [19] or as if there were a “law of small numbers” [20] according to which short sequences are more typical of long sequence mean behavior than actually predicted for a fair coin.

All of the analysis done here can be repeated in a model where “bias” is replaced with serial correlation. Then the natural parameter is not a local magnetic field but a local exchange interaction J_n between neighboring spins σ_n and σ_{n-1} . The analog of z [Eq (14)] which controls the likelihood of a sequence being assigned as random or

correlated is played by

$$z' = \frac{1}{\sqrt{4N\tau_c}} \left[\sum_{nm} (\sigma_n \sigma_{n+1}) \hat{C}_{nm} (\sigma_m \sigma_{m+1}) - N \right], \quad (24)$$

where \hat{C} is now the correlation matrix and τ_c is the correlation time for fluctuations in J . Again the sequences identified as biased will be those with large positive z' , corresponding to an excess of *either* repetitions or alternations. This may provide an even more accurate description of human perceptual biases toward representativeness in small samples [2, 19].

The model of dynamic biases considered here is closely related to the experiments in Refs [6] and [7], where animals experience time dependent biases in reward probability and have to modulate their behaviors accordingly. In these experiments the possibility of a fair coin is excluded by construction, so the question is not to determine whether the bias exists but rather to determine its current value and use this value in decision making [21]. Within the model above one can show that, for weak biases, the optimal estimate is $\hat{h}_n = \sum_{i \geq 0} K(i) \sigma_{n-i}$, where the kernel K is determined by

$$\sum_m \hat{C}_{nm} K(m-j) + \frac{1}{h_{\text{rms}}^2} K(n-j) = \hat{C}_{nj} \quad (25)$$

Specifically, if $C(\tau) = \exp(-|\tau|/\tau_c)$, then the kernel is also exponential, $K(\tau) \propto \exp(-\tau/\tau_{\text{int}})$, where the integration time is $\tau_{\text{int}} = \tau_c / \sqrt{1 + 2h_{\text{rms}}^2 \tau_c}$. Thus, when correlation times are long, even if h_{rms} is not too large the optimal estimator integrates for a time much shorter than the correlation time—e.g., with correlation times over order 100 events, $\tau_{\text{int}} \sim 10 - 15$ events, in reasonable agreement with the results of Ref [7].

In the framework considered here, the failure to recognize genuinely random sequences arises not as a limitation but rather as an inevitable consequence of the optimal search for weak dynamic biases. Important tests of this idea thus include generalizations of the experiments in Refs [6, 7] that allow detailed comparisons of human strategies for bias estimation with the optimal strategy, especially the prediction that the optimal strategy shifts with the context defined by the distribution out of which the fluctuating biases are drawn. Although not a complete theory, it seems plausible that the objective difficulty in inducing the concept of a truly fair coin found here could be related to other problems in reasoning about probability that usually are ascribed to subjective factors.

I thank GS Corrado, A Montagnini, R Rumiati, RR de Ruyter van Steveninck, SP Strong, LP Sugrue and WT Newsome for helpful discussions. Recent arguments with A Elga and D Osherson provided a stimulus to finish up; thanks to the Princeton University Council on the Humanities for the events that catalyzed these arguments.

This work was supported in part by National Science Foundation Grant IIS-0423039, through the program for Collaborative Research in Computational Neuroscience.

-
- [1] D Kahneman, P Slovic & A Tversky, eds. *Judgement Under Uncertainty: Heuristics and Biases* (Cambridge University Press, Cambridge, 1982).
 - [2] MA Bar-Hillel & WA Wagenaar, in *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, G Keren & C Lewis, eds., pp. 369–393 (Erlbaum, Hillsdale NJ, (1993); RS Nickerson, *Psych Rev* **109**, 330–357 (2002).
 - [3] A classic example is the widespread belief that basketball players can have a “hot hand;” see T Gilovich, R Vallone & A Tversky, *Cognitive Psych* **17**, 295–314 (1985).
 - [4] DM Green & JA Swets, *Signal Detection Theory and Psychophysics* (Wiley, New York, 1966).
 - [5] RHS Carpenter & MLL Williams, *Nature* **377**, 59–62 (1995).
 - [6] CR Gallistel, TA Mark, AP King & PE Latham, *J Exp Psych: Animal Behav* **27**, 354–372 (2001).
 - [7] LP Sugrue, GS Corrado & WT Newsome, *Science* **304**, 1782–1787 (2004).
 - [8] HB Barlow, *Proc R Soc Lond Ser B* **212**, 1–34 (1981).
 - [9] W Bialek, *Ann Rev Biophys Biophys Chem* **16**, 455–478 (1987).
 - [10] W Bialek, in *Physics of Biomolecules and Cells: Les Houches Session LXXV*, H Flyvbjerg, F Jülicher, P Ormos & F David, eds, pp 485–577 (EDP Sciences, Les Ulis; Springer-Verlag, Berlin, 2002). Also available at <http://arXiv.org/abs/physics/0205030>.
 - [11] CM Harris & DM Wolpert, *Nature* **394**, 780–784 (1988).
 - [12] KP Kording & DM Wolpert, *Nature* **427**, 244–247 (2004).
 - [13] J Torrmehäuser, S Gephstein, LT Maloney, MS Landy & MS Banks, *J Neurosci* **25**, 7169–7178 (2005).
 - [14] LL Lopes & GC Oden, *J Exp Psych: Learning, Memory, and Cognition* **13**, 392–400 (1987).
 - [15] Although this is a natural formulation for physicists, it is worth emphasizing any bias parameter can always be rewritten as an effective magnetic field in this notation.
 - [16] If $P(h_{\text{rms}})$ vanishes as $\sim h_{\text{rms}}^\alpha$ (or diverges for $\alpha < 0$), then in place of Eq (23) we will have $P_{\text{error}} \propto N^{-(1+\alpha)/4}$.
 - [17] Some care is required to be sure that the central limit behavior is actually valid in the regime of z relevant for computing P_{error} . The key point is that z_+ grows more slowly than any power of N .
 - [18] The optimal error probability could be even larger if the cost of missing ordered behavior is greater than the cost of responding to spurious coincidences.
 - [19] D Kahneman & A Tversky, *Cognitive Psych* **3**, 430–454 (1972).
 - [20] A Tversky & D Kahneman, *Psych Bull* **76**, 105–110 (1971).
 - [21] Strictly speaking we can pose the problem only in terms of the decisions, but it seems plausible that at least an implicit estimate of the current bias will be a step in the process of decision making. See [7] and forthcoming work from the same authors.