Piecewise Versus Total Support:

How to Deal with Background Information in Likelihood Arguments

Benjamin C. Jantzen

Virginia Tech

The probabilistic notion of *likelihood* offers a systematic means of assessing “the relative merits of rival hypotheses in the light of observational or experimental data that bear upon them.”[1] In particular, likelihood allows one to adjudicate among competing hypotheses by way of a two-part principle:

Law of Likelihood (LL):[2]

(i) Evidence *E* supports hypothesis H_1 over H_2 just if $P(E|H_1) > P(E|H_2)$, where $P(E|H_i)$ is the likelihood of hypothesis H_i given evidence *E*.

(ii) The degree to which *E* supports H_1 over H_2 is measured by the *likelihood ratio*, $\Lambda = \frac{P(E|H_1)}{P(E|H_2)}$.

The claims sanctioned by LL are strictly comparative. The principle does not say what you should believe or to what degree you should believe it. Rather, the notion of ‘supporting’ one hypothesis over another is contrastive and perhaps best characterized as a relation of ‘favoring’.[3] LL tells you how to determine the degree to which one hypothesis is favored over another on the basis of some evidence, E, and nothing more. Proponents of the principle are adamant that LL

[1] A. W. F. Edwards, *Likelihood* (Cambridge: Cambridge University Press, 1972) p. 1.

[2] I am using Elliot Sober’s terminology here. The Law of Likelihood as I’ve presented it is to be distinguished from the weaker “Likelihood Principle,” which in most formulations is equivalent to part (i) of LL. I caution the reader that both the terms “Law of Likelihood” and “Likelihood Principle” are used ambiguously in the philosophy of statistics and inductive inference literature.

[3] Elliott Sober, *Evidence and Evolution: The Logic Behind the Science* (Cambridge: Cambridge University Press, 2008).

cannot provide sufficient grounds for apportioning belief, only ranking hypotheses in a particular evidentiary context.

While LL has been defended at length as a general tool for both formal and informal reasoning about hypothesis ranking,[4] there remains an important ambiguity its application. Intuitively, we ought to make use of all available information when assessing the relative merits of two hypotheses, not just the particular piece of evidence *E* under consideration. Any additional information already in our possession prior to obtaining *E* is typically referred to as *background information*. LL does not, on the face of it, tell us how to deal with such information. Some, most prominently Elliott Sober[5], have argued that we ought to condition on this additional information when computing likelihoods. That is, if we denote the background information by *B*, then the likelihood ratio we should use is $\Lambda = \frac{P(E|H_1,B)}{P(E|H_2,B)}$. Taking this approach, however, means that Λ—and thus our judgments concerning rival hypotheses H_1 and H_2—will depend on exactly which information is taken to constitute background information, and which is considered evidence and thus part of *E*. Under Sober's interpretation, LL can be taken to yield different judgments for the

[4] See, for instance, Edwards, *Likelihood*; Ian Hacking, *Logic of Statistical Inference* (Cambridge: Cambridge University Press, 1965); Sober, *Evidence and Evolution: The Logic Behind the Science*.

[5] Elliott Sober, "The Design Argument," *God and Design*, ed. Neil Manson (New York, NY: Routledge, 2003) 27-54; Sober, *Evidence and Evolution: The Logic Behind the Science*; Elliott Sober, "Absence of Evidence and Evidence of Absence: Evidential Transitivity in Connection with Fossils, Fishing, Fine-Tuning, and Firing Squads," *Philosophical Studies* 143 (2009): 63-90.

same data when the line between evidence and background information is moved. The use of LL is thus encumbered by a “line-drawing problem.”[6]

This line-drawing problem also appears in a slightly different guise in the literature on statistical inference. In this more restricted context, the problem manifests as an apparent ambiguity in the likelihood function. Specifically, there appears to be no systematic way of deciding which random variables and model parameters should be included in the likelihood function, and no principled way of deciding on which side of the conditionalization bar these quantities belong if included.[7] As in the general case, the problem for the likelihoodist is to provide a principled division of propositions into background and evidence.

A variety of solutions have been proposed to both versions of the problem of background information, though not always in these terms. Some, e.g. Jonathan Weisberg,[8] attempt to provide a principled means of distinguishing evidence from background information. Others, e.g. Matthew Kotzen,[9] attempt to dissolve the problem by scrapping LL. In the context of statistical

[6] M. Kotzen, "Selection Biases in Likelihood Arguments," *The British journal for the philosophy of science* (2012).

[7] See M. J. Bayarri, M. H. DeGroot and J. B. Kadane, "What Is the Likelihood Function?," *Statistical Decision Theory and Related Topics Iv*, eds. Shanti S. Gupta and James O. Berger, vol. 1 (New York: Springer-Verlag, 1987) 3-27.

[8] Jonathan Weisberg, "Firing Squads and Fine-Tuning: Sober on the Design Argument," *British Journal for the Philosophy of Science* 56 (2005): 809-21.

[9] Kotzen, "Selection Biases in Likelihood Arguments."

inference, a common strategy is to disambiguate the likelihood function by fiat.[10] I argue that none of these strategies is well-motivated. Background information is only problematic when one fails to distinguish between two related questions: (i) Given that I know *B*, to what degree does the additional piece of evidence *E* support H_1 over H_2? and (ii) to what degree does all the evidence to hand—*B* and *E*—support H_1 over H_2? My aim is to demonstrate that, once these questions are distinguished the very same considerations that motivate the adoption of LL entail distinct answers to both questions, thus resolving any ambiguity over the treatment of background information. Note that I am emphatically not offering a defense of LL as a general inference procedure. Mine is the more modest goal of dissolving an apparent defect of LL using the resources to which proponents of the principle already assent.

To draw out the distinction relevant to eliminating the problem of background information, I will begin with a detailed example. I will then argue for an expression that represents the degree to which a particular piece of evidence supports one hypothesis over another in context, and then derive a related expression for the total support provided by all available evidence. Finally, I will show how these new expressions dissolve ambiguities in the treatment of background information by applying them to the so-called ‘fine-tuning argument’.

I. ILLUSTRATING THE PROBLEM

[10] See, e.g., Jason Grossman, "The Likelihood Principle," *Philosophy of Statistics*, eds. Malcolm R. Forster and Prasanta S. Bandyopadhyay (Oxford, UK; Burlington, MA: North-Holland, 2011).

To draw out the distinction which I claim obviates the problem of background information, it will help to have a concrete example in mind. To avoid pre-conceived interpretations, I will intentionally eschew standard examples, at least at the outset. So rather than treat of fish or firing squads, I'll consider carnivals.

Suppose that Albert finds himself on the midway of an old-fashioned carnival. He decides to play one of the games—the one where contestants try to toss a ball into a milk-can. Albert is savvy about carnival games; he knows they are often rigged. In a fair game, there is a 50% chance of winning a prize. But when no authorities are around, there is an appreciable chance that the carnie running the game will hand him a ball too big to fit in the can, making it impossible to win. On the other hand, if there happens to be a police officer in sight the game is likely to be rigged in Albert's favor—the carnies want the police to think the games are fair, so they arrange to let people win when the authorities are present. A set of probabilities reflecting these facts is provided by the joint distribution of Table 1.

Table 1.

	P = police present		P = police absent	
	G = fair	G = rigged	G = fair	G = rigged
O = lose	1/20	1/20	1/10	11/20
O = win	1/20	1/10	1/10	0

Knowing all of the probabilities in Table 1, Albert puts his money down, and promptly tosses a ball into the can. Given that he has just won, what can Albert conclude about the game? Specifically, does he now have grounds to favor the hypothesis that the game is fair over the hypothesis that it is rigged? According to LL, Albert needs to compare two probabilities: the probability that he would win given that the game is fair, $P(\text{win}|\text{ fair})$ and the probability that he would win given that the game is rigged $P(\text{win}|\text{ rigged})$. Since $P(\text{win }|\text{ fair}) = 1/2 > P(\text{win }|\text{ rigged}) = 1/7$, LL asserts that Albert's success in the game supports the hypothesis of a fair game—Albert has reason to think that he has played a fair game.

But suppose that, before he tosses the ball, Albert notices a police officer standing near the booth. What can be said in light of this additional information? Here is where different interpretations of LL begin to diverge. According to Sober's approach, we must recognize two sorts of propositions: evidence and background knowledge. Evidence is whatever fresh information we are currently considering when applying LL to distinguish among hypotheses. It appears to the left of the conditionalization bar when computing a likelihood. Background knowledge constitutes whatever we already know about the world, and is presumed to belong on the right side of the conditionalization bar. According to this view then, Albert should treat the fact of the police officer's presence as background knowledge and condition on this information. The relevant likelihoods are now $P(\text{win}|\text{ fair, present}) = 1/2$ and $P(\text{win}|\text{ rigged, present}) = 2/3$. With the additional information, he should now favor the hypothesis that the game is rigged—the background information has reversed our ordering on hypotheses.

That we should take all available information into account when comparing hypotheses is not especially controversial—most authors assume some sort of *principle of total evidence*.[11] What is controversial is how and whether ‘evidence’ should be distinguished from background information. It is not clear why Albert should treat the information that a police officer was present any differently than the information that he won the game. Albert might just have well have treated the observation of the police officer as the evidence, and conditioned instead on the fact that he won: $P(\text{present} \mid \text{rigged, win}) = 1 > P(\text{present} \mid \text{fair, win}) = 1/3$. In this way of accounting for all the information, LL still favors the hypothesis that the game is rigged, but does so to a much greater degree. Alternatively, Albert might have treated all the information at hand as ‘evidence’ and compared the following likelihoods: $P(\text{win, present} \mid \text{fair}) = 1/6 > P(\text{win, present} \mid \text{rigged}) = 1/7$. Taking this approach once again inverts the ordering of hypotheses, and favors the hypothesis that the game was fair. It might appear then that LL must be modified in order to provide a principled means of discriminating background information from evidence. However, no such modification is required—a careful interpretation of LL as it stands obviates the question of evidence versus background information.

II. THE PIECEWISE IMPACT OF EVIDENCE

To resolve the ambiguity over background information, we need to distinguish between two questions: (i) to what degree does learning a particular fact in the context of an additional set of facts support a given hypothesis? and (ii) to what degree does learning a particular fact in conjunction with an additional set of facts support a given hypothesis? In terms of the midway

[11] Rudolph Carnap, "On the Application of Inductive Logic," *Philosophy and Phenomenological Research* 8, 1 (1947): 133-48.

example above, the distinction can be made as follows: (i) to what degree does winning the game having already learned that a police officer is present support the hypothesis that the game is fair? and (ii) to what degree does the full set of information at hand—that Albert has won the game and that a police officer was present—support the hypothesis that the game is fair?

To address question (i), we need to examine the piecewise introduction of evidence, taking care to note one important fact: learning the truth of a proposition (or the value of a random variable) is effectively an intervention that changes the background distribution describing the ways the world might be. To begin with, let's assume that we are given a full joint distribution reflecting all relevant aspects of the world and nothing else—there is nothing given that might qualify as either evidence or background information. For ease of exposition, I will further assume that this distribution is discrete, though nothing about my derivation hinges on this being the case.

Since all we have is the distribution and no information to sort out, LL can be applied unambiguously upon obtaining our first piece of evidence, I_1. According to LL, the degree to which this information supports hypothesis H_1 over H_2 is given by the likelihood ratio $\Lambda(I_1) = P(I_1|H_1)/P(I_1|H_2)$. Furthermore, on learning that I_1 is the case, the space of possible events has been reduced—acquiring information requires us to update the background distribution with which we started. Specifically, the probability of I_1 being the case must now be unity, irrespective of the value it had prior to learning this outcome. One way to represent the change is to construct a new event space by simply removing all the events incompatible with the fact that I_1 is the case while preserving the relative measure on all remaining events. That is, the new distribution $P_1(\alpha)$, where α is any event in the original event space compatible with I_1, is obtained from the old distribution by the following relation: $P_1(\alpha) = P(\alpha|I_1)$.[12] In the midway example, for instance, when Albert learned that a police officer was present he should have replaced the original distribution of Table 1 with that of Table 2.

Table 2.

	P = police present	
	G = fair	G = rigged
O = lose	1/5	1/5
O = win	1/5	2/5

Once we realize that we are working with a new distribution, there is no need to draw a line between background information and evidence—our prior information is reflected in the new distribution. When additional evidence, I_2, is acquired, we need only appeal to LL just as we did at the outset. This time, however, we are assessing likelihoods with respect to the currently applicable distribution $P_1(\alpha)$. So the evidence I_2, if we take LL seriously, supports H_1 over H_2 just if $P_1(I_2|H_1) > P_1(I_2|H_2)$ and does so to a degree $\Lambda(\mathrm{I}_2) = P_1(I_2|H_1)/P_1(I_2|H_2)$. In terms of the original joint distribution, we can express this likelihood ratio as

$\Lambda(\mathrm{I}_2) = P(I_2|I_1, H_1)/P(I_2|I_1, H_2)$.

[12] This is simply the updating procedure recommended by Bayesian epistemology. It is invoked here without any commitment to the subjective or objective status of priors.

As before, when we learn I_2, we must update our distribution to reflect this restriction of the possibilities. This new distribution $P_2(\beta)$ is obtained from the old distribution in the same way as above: $P_2(\beta) = P_1(\beta|I_2) = P(\beta|I_2, I_1)$. This is easy to generalize for an indefinite sequence of evidence: once we've learned I_1, I_2, …, $I_{n\text{-}1}$, we should compute the likelihoods involving a new piece of evidence I_n using the distribution $P_{n-1}(\gamma) = P(\gamma|I_{n-1}, \dots, I_1)$. The new piece of information I_n introduced in the context of prior information I_1, I_2,…, $I_{n\text{-}1}$ supports H_1 over H_2 just if $P(I_n|I_{n-1}, \dots, I_1, H_1) > P(I_n|I_{n-1}, \dots, I_1, H_2)$ and does so to the degree

$$(1) \qquad \Lambda(I_n) = \frac{P(I_n|I_{n-1},\dots,I_1,H_1)}{P(I_n|I_{n-1},\dots,I_1,H_2)}.$$

The point is that whenever we acquire a piece of information we can apply LL without modification, but must do so using a distribution that reflects all of the facts already in evidence. Put this way, there is no ambiguity in using LL—we always compute a straightforward likelihood. However, when this likelihood is expressed in terms of the original joint distribution with which we started, each successive likelihood is conditioned on the previous facts. So by applying LL and taking care to note the way in which the acquisition of information forces a change in distribution, we have found that in order to determine the relative support of one hypothesis over another provided some particular piece of evidence, we must use likelihoods conditioned on all previously acquired facts.

Thus far, it may seem that I have been arguing for Sober's interpretation of LL. However, Sober seems to view the likelihood ratio (1) as representing the overall degree to which H_1 is supported over H_2 once I_n is obtained. I have been urging that, if we take LL at face value, this is *not* how we should interpret this expression. At every stage in the above derivation, we were applying LL to determine the degree to which a particular piece of evidence supported one hypothesis over another. Other information was relevant, but only in determining the epistemic context in which this degree of support was determined. I am suggesting that Sober has the right expression but gives it in answer to the wrong question—in what follows, I'll show that LL leads us to a very different expression for the degree of support for H_1 over H_2 provided by the totality of evidence.

III. TOTAL SUPPORT

There are two ways to argue for an expression of the likelihood ratio pertaining to the totality of available evidence. In one approach, we could take the expression given in (1) for the degree to which a particular piece of evidence supports H_1 over H_2 and couple this with a function for combining likelihood ratios—a function measuring the overall degree to which two pieces of evidence support H_1 over H_2. Strictly speaking, this means adding to LL since the principle does not provide such a rule. However, there are some reasonable constraints we can put on such a function without begging the question concerning background information. For starters, whatever function *f* we choose should itself yield a likelihood ratio, meaning that it must map pairs of likelihoods to the interval $[0, \infty)$. Furthermore, if either likelihood in the combination is zero—implying that one hypothesis has been entirely ruled out—then the joint likelihood should also be zero. The function should be symmetric since it ought not to matter in what order we give the likelihoods to be combined, and it should be an increasing function of both arguments. An obvious choice satisfying all of these constraints is simply the product of the component likelihoods. That is, given Λ_1 and Λ_2, the combined likelihood is given by $f(\Lambda_1, \Lambda_2) = \Lambda_1\Lambda_2$. With this rule for combining likelihoods, we can use the results of the last section to derive an expression for the overall degree to which the facts I_1, I_2, …, I_n support one hypothesis over another, assuming they were learned in sequence:

(2) $$\Lambda(I_1, I_2, \ldots, I_n) = \Lambda(I_1)\Lambda(I_2)\cdots\Lambda(I_n) = \frac{P(I_1|H_1)P(I_2|H_1,I_1)\cdots P(I_n|H_1,I_1,\ldots,I_{n-1})}{P(I_1|H_2)P(I_2|H_2,I_1)\cdots P(I_n|H_2,I_1,\ldots,I_{n-1})}$$

Using nothing but the rules of probability, the right hand side of equation (2) can be written much more compactly to give the following expression for the total support of the facts I_1, I_2, …, I_n:

(3) $$\Lambda(I_1, I_2, \ldots, I_n) = \frac{P(\,I_1,\ldots,I_n|H_1)}{P(\,I_1,\ldots,I_n|H_2)}$$

Of course, the right-hand side of equation (3) is just the expression we would have gotten by applying LL to the proposition I_1^I_2^…^I_n with respect to the initial joint distribution—in a straightforward reading, it is just the total support for H_1 over H_2 provided by the conjunction of all available evidence.

The form of Equation (3) suggests that it might have been derived more directly by appealing to LL without worrying about how to determine the contextual support provided by each piece of information or introducing a way to combine these (thus justifying my claim that we need not modify LL). All we had to do was note that, if we let $E = I_1$^I_2^ ... ^I_n, then LL immediately yields (3). From (3) we could then deduce (2) just from the rules of the probability calculus. Once we identified the factors of the right-hand side of Equation (2) with individual likelihood ratios, we could have used this fact to justify a rule for combining likelihoods. In fact, this is what A. F. Edwards does, at least in the special case of independent evidence, in his development of the likelihood framework.[13] Viewed from this perspective, Equation (3) is implicit in LL.

[13] Edwards, *Likelihood*.

Whichever approach we take to justifying this rule for assessing total support, we are led to the following amplified form of LL:

Amplified Law of Likelihoods (ALL):

(i) If it is already known to be that case that I_1^I_2^…^ I_n, then learning evidence E supports hypothesis H_1 over H_2 just if $P(E|H_1, I_1, I_2, \ldots, I_n) > P(E|H_2, I_1, I_2, \ldots, I_n)$, where $P(E|H_i, I_1, I_2, \ldots, I_n)$ is the likelihood of hypothesis H_i given evidence E in the context of I_1^I_2^…^ I_n.

(ii) The degree to which E supports H_1 over H_2 in the context of I_1^I_2^…^ I_n is measured by the likelihood ratio $\Lambda = \frac{P(E|H_1, I_1, I_2, \ldots, I_n)}{P(E|H_2, I_1, I_2, \ldots, I_n)}$.

(iii) The total evidence E^ I_1^I_2^…^ I_n supports hypothesis H_1 over H_2 just if $P(E, I_1, I_2, \ldots, I_n|H_1) > P(E, I_1, I_2, \ldots, I_n|H_2)$.

(iv) The degree to which the total evidence E^ I_1^I_2^…^ I_n supports H_1 over H_2 is measured by the likelihood ratio $\Lambda = \frac{P(E, I_1, I_2, \ldots, I_n|H_1)}{P(E, I_1, I_2, \ldots, I_n|H_2)}$.

With ALL, we can answer the questions posed above concerning the midway example. The information that Albert has won the game, acquired after learning that a police officer is present, supports the hypothesis that the game is rigged because $P(\text{win}|\text{present, rigged}) > P(\text{win}|\text{present, fair})$. According to ALL (ii), this information favors the rigged hypothesis over its rival to a degree $\Lambda = \frac{P(\text{win}|\text{present, rigged})}{P(\text{win}|\text{present, fair})} = \frac{\frac{2}{3}}{\frac{1}{2}} = \frac{4}{3}$. This one piece of information, in the context of previously established information about the presence of police officers, tends to favor the hypothesis of a rigged game. However, the aggregate information—that a police officer is present and Albert has won the game—favors the hypothesis that the game is fair. This follows from ALL (iii) and (iv) since $\frac{P(\text{win, present}|\text{ fair})}{P(\text{win, present}|\text{ rigged})} = \frac{\frac{1}{6}}{\frac{1}{7}} = \frac{7}{6}$. This looks like a contradiction until we realize that the first piece of information obtained—that the police officer is present—strongly favored the hypothesis that the game is fair: $\frac{P(\text{present}|\text{fair})}{P(\text{present}|\text{rigged})} = \frac{14}{9}$. The upshot is that the aggregate effect of the totality of evidence can differ from the piecewise impact of each bit of evidence. Rather than being a contradiction, this is precisely how one would expect these two distinct measures to relate—the total support for the fair hypothesis is simply the product of the contextual likelihood ratios for each piece of evidence.[14]

IV. KICKING AWAY THE FULL DISTRIBUTION LADDER

In the preceding arguments, I made extensive use of full probability distributions. This appears problematic since the appealing feature of the likelihood approach—and that which sets it apart from Bayesianism—is its disregard for prior probabilities. However, I claim that the likelihoodist who thinks that prior probabilities are often absent or unattainable might nonetheless justify LL or ALL. To see how, let's reconsider the case in which we start with a full prior distribution $P(\alpha)$, and then obtain evidence I_1. Once we acquire the evidence, we should update the probabilities assigned to H_1 and H_2 by setting each new probability equal to the corresponding conditional probability assigned by the original distribution:

[14] It should be noted that, while the order in which information is learned determines the degree to which each additional piece of information favors one hypothesis over another, order is irrelevant when considering the overall support conferred by the totality of evidence.

$$P\left(H_1 \middle| I_1\right) = P\left(I_1 \middle| H_1\right)\frac{P\left(H_1\right)}{P\left(I_1\right)},\ P\left(H_2 \middle| I_1\right) = P\left(I_1 \middle| H_2\right)\frac{P\left(H_2\right)}{P\left(I_1\right)}$$

What can we now say about the degree to which I_1 favors H_1 over H_2? One way we might understand this question is in terms of a hypothetical. Suppose that either H_1 or H_2 is true. Then the initial odds in favor of H_1 are simply $P(H_1|H_1 \vee H_2)/P(\sim H_1|H_1 \vee H_2) = P(H_1)/P(H_2)$. How does the new information change the odds in favor of H_1? In this case, the posterior odds are given by:

(4) $$\frac{P\left(H_1 \middle| I_1, H_1 \vee H_2\right)}{P\left(\sim H_1 \middle| I_1, H_1 \vee H_2\right)} = \frac{P\left(I_1 \middle| H_1\right)P\left(H_1\right)}{P\left(I_1 \middle| H_2\right)P\left(H_2\right)} = \Lambda\left(I_1\right)\frac{P\left(H_1\right)}{P\left(H_2\right)}$$

The right-hand equality in Equation (4) indicates that all of the work to shift the posterior odds up or down relative to our prior odds is being done by the likelihood ratio, $\Lambda(I_1)$. In other words, the change in posterior odds is a function of $\Lambda(I_1)$. To put it still another way, the effect of I_1 on the odds is entirely determined by $\Lambda(I_1)$. This fact motivates adopting the likelihood function as a measure of relative support. While the likelihood ratio cannot tell us which posterior probability is higher, it can tell us how the odds shift in favor of one hypothesis or the other, assuming that one or the other is right. Furthermore, it does so whether or not we know the prior probabilities. In this sense, LL is a general guide to differential support, and in those cases in which we have no objective basis for assigning priors, the likelihoodist claims it is our only guide.

By considering effects on posterior odds, we can motivate ALL in much the same way as LL. As before, the full distribution (if we knew it to begin with) after learning I_1 would be given by $P_1(\alpha) = P(\alpha|I_1)$. If we now learn that I_2 is the case, then we must change our posterior odds in favor of H_1 over H_2 to the following:

(5) $$\frac{P_1\left(H_1 \middle| I_2, H_1 \vee H_2\right)}{P_1\left(H_2 \middle| I_2, H_1 \vee H_2\right)} = \frac{P_1\left(I_2 \middle| H_1\right) P_1\left(H_1\right)}{P_1\left(I_2 \middle| H_2\right) P_1\left(H_2\right)} = \Lambda\left(I_2\right) \frac{P_1\left(H_1\right)}{P_1\left(H_2\right)}$$

Once again, it is the likelihood function that increases (or decreases) the posterior over the prior odds. This time, however, it is in the context of the new distribution $P_1(\alpha)$, a distribution reflecting prior knowledge of I_1. If the motivation offered for LL in the first place is compelling, then it seems we must also accept ALL (i) and (ii)—the relative support for H_1 over H_2 conferred by the new piece of evidence I_2 after already learning I_1 is indicated by the likelihood function, $\Lambda(I_2) = P_1(I_2|H_1)/P_1(I_2|H_2) = P(I_2|I_1, H_1)/P(I_2|I_1,H_2)$. But what about the overall support for H_1 over H_2 given our epistemic starting point? How should our posterior odds have changed relative to our initial odds as a result of learning I_1 and I_2? We can rewrite the right-hand side of Equation (5) as follows:

(6) $$\begin{aligned} \frac{P_1\left(I_2 \middle| H_1\right) P_1\left(H_1\right)}{P_1\left(I_2 \middle| H_2\right) P_1\left(H_2\right)} &= \frac{P\left(I_2 \middle| H_1, I_1\right) P\left(H_1 \middle| I_1\right)}{P\left(I_2 \middle| H_2, I_1\right) P\left(H_2 \middle| I_1\right)} \\ &= \frac{P\left(I_1, I_2 \middle| H_1\right) P\left(H_1\right)}{P\left(I_1, I_2 \middle| H_2\right) P\left(H_2\right)} \\ &= \Lambda\left(I_1, I_2\right) \frac{P\left(H_1\right)}{P\left(H_2\right)} \end{aligned}$$

Written this way, we can see that the combined likelihood function $\Lambda(I_1, I_2)$ determines the change in odds relative to what they were before learning anything at all. So once again, we can kick away the ladder of the full distribution. If we did know the distribution, then learning I_1 and I_2 would change the odds in favor of H_1 over H_2 by an amount given by the likelihood ratio. Since this is true irrespective of what the priors are, we can always take the likelihood alone to indicate differential support, in this case the degree of support for H_1 over H_2 conferred by the totality of evidence.

Of course, one might object to my interpretation of what it is to favor one hypothesis over another. Instead, one might attempt to prove LL(i) from other premises[15] and take the quantitative measure of contrastive support given by LL (ii) to be a postulate that stands or falls with how well the results coincide with our intuitions.[16] ALL could then be motivated by the second line of argument I suggested in the previous section: treat all information as evidence and note that the resulting likelihood ratio factors into a product of likelihoods, each of which can be consistently interpreted as corresponding to the impact of a single piece of information.

The point is that insofar as LL is well-motivated, so too is ALL. My use of full distributions above was strictly heuristic. Once we've seen what role the likelihood function plays and which likelihood function is relevant to which question, we can ignore the full distribution. Of course, the proponent of LL or ALL can only claim to be free of worrisome priors if conditional probabilities can be taken as primitive.[17] It is not my task here to defend that claim and thus rescue likelihoodism from the charge of subjectivity. My more modest assertion is simply that if we have grounds to take LL seriously, then we should really embrace ALL. Once we've done so, it becomes clear that whatever problems likelihoodism has, line-drawing isn't one of them.

V. BUT WHAT IS THE LIKELIHOOD FUNCTION?

[15] See Grossman, "The Likelihood Principle."

[16] See Sober, *Evidence and Evolution: The Logic Behind the Science*.

[17] For a defense of taking conditional probabilities as primitives, see ibid.

My arguments so far have concerned the problem of background information as it appears in the literature on LL in its broadest epistemic use. As I mentioned above, the same problem arises in the more restricted context of statistical inference. Addressing this narrower community, Bayarri, De Groot, and Kadane famously asked, “What is the likelihood function?”[18] To illustrate the ambiguity in answering that question, the authors consider a case analogous to one in which, with respect to each possible value of some discrete parameter θ characterizing a statistical model, a random variable X has a conditional probability distribution $P(x|\theta)$.[19] Furthermore, it is not the random variable X that is observed, but rather some other random variable Y for which $\mathrm{P}(y|x, \theta)$. The authors then ask, “What is the [likelihood function] in this problem?”.[20] They claim that there are three candidates, $P(y|\theta)$, $\mathrm{P}(x, y|\theta)$, and $\mathrm{P}(y|x, \theta)$, and that a “subjective judgment must be made in order to decide which of the functions…to use in a given problem.”[21] The thesis I’ve been defending is that this is simply false. The question has two parts: (i) which random variables and parameters are to be included in the likelihood function, and (ii) which side of the conditionalization bar each belongs on. The answer to both parts, according to ALL, depends on two things: what hypotheses we wish to consider and whether we wish to assess the impact of a particular piece of data in context or the aggregate of all data. So, for instance, suppose we wish to ask about hypotheses concerning the value of θ in light of the only piece of

[18] Bayarri, DeGroot and Kadane, "What Is the Likelihood Function?."

[19] The original example was stated in terms of probability densities since θ typically takes a continuum of values. To keep the discussion consistent, I’ve assumed that θ is discrete, and thus the distributions in question are discrete as well.

[20] Bayarri, DeGroot and Kadane, "What Is the Likelihood Function?," at p. 6.

[21] Ibid., 6.

evidence available, namely a value y of Y. Then the relevant likelihood function must have the form $P(y|\theta)$. If on the other hand, we wanted to consider finer-grained hypotheses concerning both the value of θ and the unobserved random variable X, then we would have functions of the form $P(y|x, \theta)$. Under no circumstances would ALL entail the use of a likelihood function of the form $P(x|\ldots)$ unless a value of the random variable X was observed (or otherwise learned) and thus added to our store of facts. Suppose X and Y were both observed and we wish to know the relative support given to hypotheses about θ. Then our likelihood functions would look like $P(x,y|\theta)$. Suppose instead, we learned the value of Y and then the value of X and wish to know what impact learning $X = x$ has given what we already know about Y. Then the likelihood functions would have the form $P(x|y, \theta)$. I'm belaboring the point, but I want to make it clear that ALL unambiguously selects a set of variables and parameter values and distributes these around the conditionalization bar. There are many further objections raised by Bayarri et al to the use of LL as a statistical inference method, in particular problems with prediction. However, many of these objections conflate LL (or ALL) with the method of maximum likelihood estimation (MLE). A discussion of the relation of MLE to ALL is beyond the scope of this paper, and so too are the remaining objections to likelihoodism. It suffices here to note that there is no ambiguity in factoring the likelihood function as far as ALL is concerned. The principle may not be right, but it is unambiguous.[22]

[22] The authors might object that in my initial discussion of ALL, I used a full distribution which dictates all the relevant quantities and so implicitly settles the question of which likelihood function to use. But as I argued in the last section, a full distribution is unnecessary for motivating ALL. Rather, in the likelihoodist view, specifying a question of interest specifies a

VI. FISH, FIRING SQUADS, AND FINE-TUNING

The question of how to handle background information is especially pressing in the context of the *fine-tuning argument* (FTA). The FTA attempts to establish the existence of a cosmic designer by noting that various physical constants have values within a narrow range amenable to the occurrence of carbon-based life—the laws appear 'fine-tuned' for life. For instance, had the 7.65 MeV energy level of the C^{12} nucleus been slightly lower or higher, then the process that produces carbon and the other heavy elements essential to life in the interior of stars would not have occurred.[23] Denote by E the observation that many constants occurring in physical laws take values within a comparatively narrow range that permits life to exist, and consider the following two hypotheses:

H_C: The relevant physical constants acquired their values by chance.

H_D: The relevant physical constants acquired their values by design.

The FTA is usually presented as a likelihood argument. If we appeal to LP and note that $P(E|H_D) > P(E|H_C)$, then we must conclude that the evidence favors design over chance.

A prominent objection to the fine-tuning argument notes that we have left out an important piece of information: all knowledge concerning physical constants has been acquired by carbon-based

likelihood ratio which in turn constrains what full distributions the Bayesian (or anyone else committed to using full distributions) may consider.

[23] John D. Barrow and Frank J. Tipler, *The Anthropic Cosmological Principle* (New York: Oxford University Press, 1986) pp. 252-53.

life forms.[24] Call this fact I. We must account for all available background information—so the objection goes—and so we must condition our likelihoods on I. However, since I entails E, both hypotheses have the same likelihood given the evidence: $P(E|H_D, I) = P(E|H_C, I) = 1$. Thus, the evidence cannot favor design over chance (or any other hypothesis for that matter). This objection, however, conflates the two questions with which we began and emphasizes the need for the clarification provided by ALL.

To motivate an analysis of the FTA in terms of ALL, it will help to first consider a pair of structurally similar examples endemic in the literature. The first of these, due originally to Sir Arthur Eddington,[25] asks us to think about fishing. Suppose we are confronted with the following observation:

E_f: All 10 of the fish caught in the lake today were longer than 10 inches.

For the sake of simplicity, suppose that we consider only two hypotheses that might account for this evidence:

H_{100}: All of the fish in the lake are longer than 10 inches.

H_{50}: Half of the fish in the lake are longer 10 inches.

If this was all the information we had, LP would urge us to favor H_{100} since $P(E_f|H_{100}) \gg P(E_f|H_{50})$. However, suppose we had some additional information:

$I_{>10}$: The net used has holes 10 inches wide.

[24] Sober, "The Design Argument."; Sober, "Absence of Evidence and Evidence of Absence: Evidential Transitivity in Connection with Fossils, Fishing, Fine-Tuning, and Firing Squads."

[25] A. Eddington, *The Philosophy of Physical Science* (Cambridge: Cambridge University Press, 1947).

This new information $I_{>10}$ entails E_f. Thus, if we account for this new information by conditioning on it as Sober would urge, we find that the evidence fails to distinguish between the hypotheses at all: $P(E_f | H_{100}, I_{>10}) = P(E_f | H_{50}, I_{>10}) = 1$. According to Sober, this constitutes an *Observation Selection Effect* (OSE) because the method by which the observation was obtained biased the outcome. One is faced with an OSE whenever accounting for the process by which an observation was made alters the likelihoods that determine the degree to which the observation favors one hypothesis over another. In this case, the effect is extreme.

The picture changes dramatically when we analyze this scenario using ALL. It becomes immediately obvious that the likelihoods being compared— $P(E_f | H_{100}, I_{>10})$ and $P(E_f | H_{50}, I_{>10})$—represent only the degree to which learning about the day's catch supports either H_{100} or H_{50} in the context of information about the net used. These do not represent the degree to which the aggregate evidence supports one or the other hypothesis. It is true that learning E after learning what net was used fails to further discriminate between H_{100} and H_{50}. But learning $I_{>10}$ may have already discriminated between the two, and thus, according to ALL, the aggregate information might also discriminate between the two hypotheses.

To illustrate the point, consider the joint distribution in Table 3. I've added a proposition, $I_{>0}$, which is the claim that the net used had very tiny holes capable of catching the smallest fish. With this additional possibility added, the probabilities given are compatible with all of the facts above. In particular, $P(E_f | H_{100}) = 1 \gg P(E_f | H_{50}) = .003$ and $P(E_f | H_{100}, I_{>10}) = P(E_f | H_{50}, I_{>10}) = 1$.

Table 3.

	H_{100}		H_{50}	
	$I_{>0}$	$I_{>10}$	$I_{>0}$	$I_{>10}$
E_f	.001	.002	.001	.002
$\neg E_f$	0	0	.994	0

However, we can see that learning $I_{>10}$ at the outset strongly favored the hypothesis H_{100} since $P(I_{>10}|H_{100}) = 0.67 \gg P(I_{>10}|H_{50}) = 0.002$. Likewise, according to ALL (iv), the aggregate information overwhelmingly favors H_{100} over H_{50} to a degree given by $\Lambda = P\left(E_f, I_{>10}\middle|H_{100}\right)/P\left(E_f, I_{>10}\middle|H_{50}\right) = 334$. This conclusion is not surprising given the details of the example. The distribution given in Table 3 is plausible in that those who frequently fish a particular lake are more likely to use nets with large holes if the lake contains mostly large fish—they may not know the distribution of fish in the lake, but they know what works. Whatever story one might tell to account for the particular probabilities in this case, the upshot is that if an OSE renders a particular observation irrelevant in a particular context it is still possible for the aggregate information to discriminate between hypotheses.

While Eddington's fishing example illustrates the way in which previously acquired information can deprive subsequent evidence of relevance, there is another example in the literature more closely analogous to the fine-tuning case.[26] This scenario involves firing squads. We are asked to imagine that a firing squad staffed by twelve expert marksmen takes aim at the prisoner to be executed. Each marksman fires twelve times when given the signal. When the smoke clears, we discover that the prisoner is still unharmed. Call the fact of this surprising survival E_s. In this case, we are interested in what the prisoner can infer from E_s concerning the following two hypotheses:

H_{con}: The marksmen conspired at time t_1 to spare the prisoner's life when they fired at t_2.

H_{miss}: The marksmen decided at time t_1 to shoot the prisoner when they fired at t_2 but missed by chance.

At first we might think that the prisoner has ample reason to favor H_{con} over H_{miss} since, given that these are expert marksmen, $P(E_s|H_{con}) \gg P(E_s|H_{miss})$. However, in making his analysis the prisoner left out some pertinent information about the manner in which the observation of E_s was made:

I_O: At t_3 the prisoner made the observation that he is still alive.

According to those who would single out background information, we must incorporate I_O into the likelihoods by conditioning. In this view, the prisoner suffers from an OSE and cannot distinguish between the two hypotheses at all since $P(E_s|H_{con}, I_O) = P(E_s|H_{miss}, I_O) = 1$. Because I_O entails E_s, so the argument goes, learning E_s can tell the prisoner nothing about which

[26] The scenario was introduced in John Leslie, *Universes* (London: Routledge, 1989). and elaborated in Richard Swinburne, "Arguments from the Fine-Tuning of the Universe," *Physical Cosmology and Philosophy*, ed. J. Leslie (New York: MacMillan, 1990) 160-79.

hypothesis to favor. Thus, the prisoner in the grip of a strong OSE cannot reasonably conclude there was a conspiracy to save his life.

At this point, the tight analogy with the FTA should be clear. The prisoner stands in for us carbon-based life forms. While the prisoner is attempting to assess whether design or chance is responsible for his survival, in the FTA we are attempting to infer design in the cosmos. In both cases, it has been objected that the observer suffers from an OSE that prevents discrimination between hypotheses. Supporters of the FTA invoke the firing-squad scenario because they think that our intuition strongly opposes the OSE objection—surely the prisoner can reasonably conclude that conspiracy is the better hypothesis. By analogy, they claim that we can conclude that an OSE is not a problem for the FTA.

In both cases, ALL tells us that the role of the OSE has been misinterpreted. It is true that, in the context of knowing that it was himself who made the observation, the prisoner learns nothing further by noting that he is alive. Likewise, it is the case that, knowing that all physics is done by carbon-based life forms, we learn nothing further by discovering that the constants of physical law are just right to sustain carbon-based life. Nonetheless, the aggregate information might still favor one hypothesis over the other. In the firing-squad scenario, it is eminently plausible that $P(E_s, I_O | H_{con}) \gg P(E_s, I_O | H_{miss})$. In the case of fine-tuning, it may be that $P(E, I | H_D) > P(E, I | H_C)$. This will be the case if $P(I | H_D) > P(I | H_C)$. I certainly do not wish to argue that this is in fact the case—there seem to be insurmountable difficulties in providing a well-defined measure corresponding to $P(I|H_D)$.[27] My point is just that, when one distinguishes between contextual and total support, the presence of an OSE does not prove fatal to design arguments in either the firing-squad or FTA case.

VII. CONCLUSION

Insofar as one is inclined to accept LL as a framework for inference, no modification is necessary in order to deal with background information—unpacking LL leads to ALL. The interpretive key is the discrimination of two questions, one concerning the immediate support provided by a piece of evidence in context and one concerning the overall support provided by the total set of evidence. Looked at in this way, it becomes clear that objections based on observer bias are not necessarily fatal to the FTA. It is true that we, as carbon-based life-forms, cannot use the fact that some physical constants are just right for the existence of carbon-based life to discriminate between design hypotheses and their rivals. However, it may be the case that the aggregate evidence (including the fact of our existence) might permit such discrimination. Whether this is the case must be settled on other grounds.

[27] It is not clear that the question of fine-tuning is even well-posed. There is reason to reject the strong metaphysical assumptions necessary to make the possibility of different ‘constants’ in the laws of nature meaningful or to entertain the existence of processes—whether physical or divine—that determined those constants in the past.