Design by Elimination vs. Design by Comparison

(Chapter 33 from *The Design Revolution*)

By William A. Dembski

How are design hypotheses properly inferred, simply by eliminating chance hypotheses or by comparing the likelihood of chance and design hypotheses?

Behind this question are two fundamentally different approaches about how to reason with chance hypotheses. One approach, due to Ronald Fisher, rejects a chance hypothesis provided sample data appear in a prespecified rejection region. The other, due to Thomas Bayes, rejects a chance hypothesis provided an alternative hypothesis confers a bigger probability on the data in question than the original hypothesis. In the Fisherian approach, chance hypotheses are rejected in isolation for rendering data too improbable. In the Bayesian approach, chance hypotheses are eliminated provided some other hypotheses render the data more probable. Whereas in the Fisherian approach the emphasis is on elimination, in the Bayesian approach the emphasis is on comparison. These approaches are incompatible, and the statistical community itself is deeply riven over which of these approaches to adopt as the right canon for statistical rationality. The difference reflects a deep divergence in fundamental intuitions about the nature of statistical rationality and in particular about what counts as statistical evidence.

The most influential criticism of specified complexity charges it with falling on the wrong side of this divide. Specifically, critics charge that to use specified complexity to infer design presupposes an eliminative, Fisherian approach to reasoning with chance hypotheses whereas the right approach to inferring design needs to embrace a comparative, Bayesian approach. The most prominent scholar to make this criticism is Elliott Sober. Other scholars have offered this criticism as well, and many more still have cited it as decisively refuting specified complexity as a sign of intelligence.

In responding to this criticism, let's begin with a reality check. Often when the Bayesian literature tries to justify Bayesian methods against Fisherian methods, authors are quick to note that Fisherian methods dominate the scientific world. For instance, Richard Royall (who strictly speaking is a likelihood theorist rather than a Bayesian—the distinction is not crucial to this discussion) writes: "Statistical hypothesis tests, as they are most commonly used in analyzing and reporting the results of scientific studies, do not proceed ... with a choice between two [or more] specified hypotheses being made ... [but follow] a more common procedure...." (*Statistical Evidence: A Likelihood Paradigm*, Chapman & Hall, 1997.) Royall then outlines that common procedure, which requires specifying a single chance hypothesis, using a test-statistic to identify a rejection region, checking whether the probability of that rejection region under the chance hypothesis falls below a given significance level, determining whether a sample (the data) falls within that rejection region, and if so rejecting the chance hypothesis. In other words, the sciences look to Ronald Fisher and not Thomas Bayes for their statistical methodology. Howson and Urbach, in *Scientific Reasoning: The Bayesian Approach*, likewise admit the underwhelming popularity of Bayesian methods among working scientists.

So, are the majority of scientists just being stupid or lazy in adopting a Fisherian approach to statistical reasoning? To answer this question, let's look at two prototypical examples where Fisherian and Bayesian methods are employed. Once these examples are in hand, we can tinker with them to see what can go wrong with both methods. Let's start with an example of Fisherian reasoning. The Fisherian approach eliminates chance hypotheses in isolation, so we need only consider a single chance hypothesis for elimination. Let's take a particularly simple one, namely, the chance hypothesis that characterizes the tossing of a fair coin. To test whether the coin is biased in favor of heads, and thus not fair, one can set a rejection region of ten heads in a row and then flip the coin ten times. In Fisher's approach, if the coin lands ten heads in a row, then one is justified rejecting the chance hypothesis. The improbability of tossing ten heads in a row, assuming the coin is fair, is approximately one in a thousand (i.e., .001).

Next, to illustrate the Bayesian approach, consider the following probabilistic set-up. Imagine two coins, the one fair and the other biased. Assume the biased coin has probability of landing heads ninety percent of the time. In addition, imagine a giant urn with a million equally sized balls, all of which except one are white, the lone exception being black. Now imagine that a single random sample will be taken from the urn and that if a white ball is selected (which is overwhelmingly probable), then the fair coin will be tossed ten times, but if the one lone black ball is selected (which is overwhelmingly improbable), then the biased coin will be tossed ten times. Now imagine that all you see is a coin tossed ten times and each time landing heads. The probability of it landing ten heads in a row given that the fair coin was tossed is approximately .001 (one in a thousand). But the probability of it landing ten heads in a row given that the biased coin was tossed is approximately .35 (a little better than one in three). Within the Bayesian literature, these probabilities are known as *likelihoods*.

So which coin was tossed, the fair one or the biased one? If one looks purely at likelihoods, it appears that the biased coin was tossed—indeed, it's much more likely that ten heads in a row will appear from the biased coin than from the fair coin. But that answer will not do. The problem is that which coin gets tossed has what in the Bayesian literature is called a *prior probability*. That prior probability renders it much more likely that the fair coin was tossed than the biased coin. The fair coin has prior probability .999999 of being tossed (because a white ball is that likely to be selected from the urn) whereas the biased coin has prior probability .000001 of being tossed (because the one lone black ball is only that likely to be selected from the urn).

To decide which coin was tossed, these prior probabilities need to be factored into the likelihoods calculated earlier. To do that, one calculates what in the Bayesian literature are known as *posterior probabilities* (these are calculated via Bayes's Theorem). The posterior probability for the fair coin being tossed given that ten heads in a row were observed is .9996 whereas the posterior probability for the biased coin being tossed given that ten heads in a row were observed is .0004. Given the probabilistic set-up for the two coins and urn as described above, it is therefore much more probable that the fair coin was tossed than the biased coin. And this is the case even though the observed outcome of ten heads in a row taken by itself is more consistent with the biased coin than with the fair coin.

Given these particularly neat and clean illustrations of the Fisherian and Bayesian approaches, one might wonder what's the problem with either. Both approaches, as illustrated in these examples, seem eminently reasonable given the questions they are called to answer. Nevertheless, both approaches raise serious conceptual problems when probed more deeply. I want in the remainder of this chapter to describe the conceptual problems raised by the Fisherian approach and indicate how my work on specified complexity helps resolve them. Next I want to describe the conceptual problems raised by the Bayesian approach and indicate why they render it inadequate as a general model of statistical rationality. In particular, I show how the Fisherian approach can be made logically coherent and why the Bayesian approach, when it works (which is not too often), must in fact presuppose the Fisherian approach.

So, what are the problems with the Fisherian approach and how does my work on specified complexity help resolve them? Schematically, the Fisherian approach looks as follows: A chance hypothesis defined with respect to a reference class of possibilities is given. Also given is a rejection region from that reference class. With the chance hypothesis and rejection region in place, an event is then sampled from the reference class of possibilities. If that event (the sample or data) falls within the rejection region and if the probability of that rejection region with respect to the chance hypothesis is sufficiently small, then the chance hypothesis is rejected. Intuitively, think of an arrow shot at a large wall displaying a fixed target. The wall corresponds to the reference class of possibilities (all the places the arrow might land) and the target to the rejection region. Provided that the arrow landing in the target (i.e., the sample falling in the rejection region) has sufficiently small probability, then the chance hypothesis is rejected. In our earlier coin tossing example, the rejection region was all possible sequences of heads and tails, the rejection region was all sequences beginning with ten heads in a row, the sample was a sequence of ten heads in a row, and the chance hypothesis presupposed a fair coin.

Is there something wrong with this picture? Although this picture has proven quite successful in practice, Ronald Fisher, in formulating its theoretical underpinnings, left something to be desired. There are three main worries: (1) How does one make precise what it means for a rejection region to have "sufficiently small" probability with respect to a chance hypothesis? (2) How does one characterize rejection regions so that a chance hypothesis doesn't automatically get rejected in case it actually is operating? (3) Why should a sample that falls in a rejection region count as evidence against a chance hypothesis?

The first concern is usually stated in terms of setting a "significance level." A significance level prescribes the degree improbability below which a rejection region eliminates a chance hypothesis once the sample falls within it. Significance levels in the social sciences literature, for instance, usually weigh in at .05 or .01. But where do these numbers come from? In fact, they are entirely arbitrary. This arbitrariness has dogged the Fisherian approach from the start. Nevertheless, there is a way around it.

Consider again our example of tossing a coin ten times and getting ten heads in a row. The rejection region, which matches this sequence of coin tosses, therefore sets a significance level of .001. If we tossed ten heads in a row, we might therefore regard this as evidence against the coin being fair. But what if we didn't just toss the coin ten times on one occasion but tossed it ten times on multiple occasions? If most of the time we tossed the coin its behavior was entirely

what one would expect from a fair coin, then on those few occasions when we observed ten heads in a row, we would have no reason to suspect that the coin was biased since fair coins, if tossed sufficiently often, will produce any sequence of coin tosses, including ten heads in a row. The strength of the evidence against a chance hypothesis when a sample falls within a rejection region therefore depends on how many samples are taken or might have been taken. These samples constitute what I call *replicational resources*. The more such samples, the greater the replicational resources

Significance levels therefore need to factor in replicational resources if samples that match these levels are to count as evidence against a chance hypothesis. But that's not enough. In addition to factoring in replicational resources, significance levels also need to factor in what I call *specificational resources*. The rejection region on which we've been focusing specified ten heads in a row. But surely if samples that fall within this rejection region could count as evidence against the coin being fair, then samples that fall within other rejection regions must likewise count as evidence against the coin being fair. For instance, consider the rejection region that specifies ten tails in a row. By symmetry, samples that fall within this rejection region must count as evidence against the coin being fair just as much as samples falling within the rejection region that specifies ten heads in a row.

But if that is the case, then what's to prevent the entire range of possible coin tosses from being swallowed up by rejection regions so that regardless what sequence of coin tosses is observed, it always ends up falling in some rejection region and therefore counting as evidence against the coin being fair? More generally, what's to prevent any reference class of possibilities from being partitioned into a mutually exclusive and exhaustive collection of rejection regions so that any sample will always fall in some one of these rejection regions and therefore count as evidence against any chance hypothesis whatsoever?

The way around this concern is to limit rejection regions to those that can be characterized by low complexity patterns (such a limitation has in fact been implicit when Fisherian methods are employed in practice). Rejection regions, and specifications more generally, correspond to events and therefore have an associated probability or probabilistic complexity. But rejection regions are also patterns and as such have an associated complexity that measures the degree of complication of the patterns, or what I call its specificational complexity. Typically this form of complexity corresponds to a Kolmogorov compressibility measure or minimum description length (the shorter the description, the lower the specificational complexity—see http://www.mdl-research.org). I summarize these two types of complexity in chapter 10. Note, specificational complexity arises very naturally—it is not artificial or ad hoc construct designed simply to shore up the Fisherian approach. Rather, it has been implicit right along, enabling Fisher's approach to flourish despite the inadequate theoretical underpinnings that Fisher provided for it.

Replicational and specificational resources together constitute what I call *probabilistic resources*. Probabilistic resources resolve the first two worries raised above concerning Fisher's approach to statistical reasoning. Specifically, probabilistic resources enable us to set rationally justified significance levels, and they constrain the number of specifications, thereby preventing chance hypotheses from getting eliminated willy-nilly. Probabilistic resources therefore provide

a rational foundation for the Fisherian approach to statistical reasoning. What's more, by estimating the probabilistic resources available in the known physical universe, we can set a significance level that's justified irrespective of the probabilistic resources in any given circumstance. Such a context-independent significance level is thus universally applicable and definitively answers what it means for a significance level to be "sufficiently small" regardless of circumstance. For a conservative estimate of this significance level, known as a universal probability bound, see chapter 10. For the details about placing Fisher's approach to statistical reasoning on a firm rational foundation, see chapter 2 of *No Free Lunch*.

That leaves the third worry concerning the Fisherian approach to statistical reasoning, namely, Why should a sample that falls in a rejection region (or, more generally, an outcome that matches a specification) count as evidence against a chance hypothesis? Once one allows that the Fisherian approach is logically coherent and that one can eliminate chance hypotheses individually simply by checking whether samples fall within suitable rejection regions (or, more generally, outcomes match suitable specifications), then it is a simple matter to extend this reasoning to entire families of chance hypotheses, perform an eliminative induction (see chapter 31), and thereby eliminate all relevant chance hypotheses that might explain a sample. And from there it is but a small step to infer design.

Let's stay with this last point for a moment—how does one go from eliminating chance to inferring design? Indeed, what justifies this move from chance elimination to design inference? We are supposing, for the moment, that the Fisherian approach can legitimately eliminate individual chance hypotheses and thus, by successive elimination, eliminate whole families of chance hypotheses. To eliminate a chance hypothesis, the Fisherian approach determines whether an outcome matches a specification and whether the specification itself describes an event of small probability (the event here comprises all outcomes that match the specification). Given that we've successfully characterized all chance hypotheses that exclude design and that we've been able to eliminate them by means of such a specification (the outcome therefore exhibits specified complexity), why should we think that outcome is designed?

In this case the specification itself acts as a logical bridge between chance elimination and design inference. Here's the rationale: If we can spot an independently given pattern (i.e., specification) in some observed outcome and if possible outcomes matching that pattern are, taken jointly, highly improbable (in other words, the observed outcome exhibits specified complexity), then it's more plausible that some end-directed agent or process produced the outcome by purposefully conforming it to the pattern than that it simply by chance ended up conforming to the pattern. Accordingly, even though specified complexity establishes design by means of an eliminative argument, it is not fair to say that it establishes design by means of a *purely* eliminative argument. The independently given pattern, or specification, contributes positively to our understanding of the design inherent in things that exhibit specified complexity.

To avoid this slippery slope to design, Bayesian theorists deny that the Fisherian approach can legitimately eliminate even one chance hypothesis (much less sweep the field clear of all relevant chance hypotheses as required for a successful design inference). The problem, as they see it, is that samples falling within rejection regions (or, more generally, outcomes matching specifications) cannot serve as evidence against chance hypotheses. Rather, the only way for

there to be evidence against a chance hypothesis is for there to be better evidence in favor of some other hypothesis.

I'll analyze the Bayesian approach to statistical evidence momentarily, but first I need to say a word about evidence generally. In *World Without Design*, Michael Rea remarks, "True inquiry is a process in which we try to revise our beliefs on the basis of what we take to be evidence." He continues, "But this means that, in order to inquire into anything, we must *already* be disposed to take some things as evidence. In order even to begin inquiry, we must already have various dispositions to trust at least some of our cognitive faculties as sources of evidence and to take certain kinds of experiences and arguments to be evidence. Such dispositions (let's call them *methodological dispositions*) may be reflectively and deliberately acquired."

Accordingly, what counts as evidence (and that includes statistical evidence) is decided not on the basis of evidence but on the basis of dispositions that themselves are not mandated by evidence. Why, for instance, do most mathematicians find proof by contradiction (i.e., *reductio ad absurdum*) as compelling evidence for the truth of a mathematical proposition, but others (the intuitionists) find such proofs inadequate and instead require constructive proofs? Or again, why do Fisherian and Bayesian approaches to statistical evidence remain at loggerheads? In such cases the debate is not merely over how to weigh certain evidence but over what counts as evidence in the first place. The issue of what counts as evidence cuts across the entire debate over intelligent design. Can there even be such a thing as evidence for an unevolved intelligence that designs biological complexity? Many naturalistic scientists and philosophers deny it. But to deny it coherently, one needs an evidential framework for denying it. The preeminent framework in that regard is Bayesian. I want therefore next to examine that framework and specifically to show why it is inadequate both for drawing design inferences as well as for precluding them.

When the Bayesian approach tries to adjudicate between chance and design hypotheses, it treats both chance and design hypotheses as having prior probabilities and as conferring probabilities on outcomes and events. Thus, given the chance hypothesis H, the design hypothesis D, and the outcome E, the Bayesian theorist attempts to compare the posterior probabilities of H and D on E (i.e., P(H|E) vs. P(D|E)). If the posterior probability of D on E is greater than that of H on E, then E counts as evidence in favor of D, and the strength of that evidence is proportional to how much greater P(D|E) is than P(H|E). Unfortunately, calculating posterior probabilities requires knowing prior probabilities (i.e., P(H) and P(D)), and often these are not available. In that case, one may merely calculate the likelihoods of E on both H and D (i.e., P(E|H) vs. P(E|D)).

There's a stripped down version of the Bayesian approach known as the likelihood approach that essentially ignores prior probabilities and simply looks at the likelihood ratio (i.e., P(E|H)/P(E|D)) to determine strength of evidence in favor of a hypothesis. This, however, makes for an idiosyncratic understanding of evidence. Evidence, as usually understood, refers to what causes us to revise our beliefs. But likelihoods ratios are in no position to do that without help from prior probabilities. For instance, if I hear from my attic the pitter-patter of little feet and the sound of bowling pins colliding, the likelihood of the design hypothesis that gremlins are bowling in my attic may be greater than the likelihood of any chance hypothesis that purports to explain those sounds. And yet, my disbelief in the gremlin hypothesis would remain as utter and

complete as before because of my prior belief that gremlins don't exist (in Bayesian terms, the prior probability P(D), where D is the gremlin hypothesis, is for me effectively zero).

I've just described the Bayesian approach to assessing the evidence for design hypotheses in comparison with chance hypotheses. Accordingly, to draw a design inference is to determine that the evidence, construed in Bayesian or likelihood terms, favors design over chance. What's wrong with this approach inferring design? Lots. I'll briefly summarize what's wrong bullet-point fashion. For more details, refer to chapter 2 of *No Free Lunch*.

(1) Need for prior probabilities. As we've already seen, for the Bayesian approach to work requires prior probabilities. Yet prior probabilities are often impossible to justify. Unlike the example of the urn and two coins discussed earlier, in which drawing a ball from an urn neatly determines the prior probabilities regarding which coin will be tossed, for most design inferences, especially the interesting ones like whether there is design in biological systems, we have no handle on the prior probability of a design hypothesis, or that prior probability is fiercely disputed (theists, for instance, might regard the prior probability as high whereas atheists would regard it as low).

(2) Design hypotheses conferring probabilities. The Bayesian approach requires that design hypotheses, as with chance hypotheses, confer probabilities on events. In the notation above, for the Bayesian approach to work, the likelihoods P(E|D) and P(E|H) both need to be well-defined. Suppose E denotes the event responsible for a certain gene, where this gene in turn codes for a certain enzyme. Given the various natural processes to which genes are subject (mutation, deletion, duplication, cross-over, etc.), P(E|H) is well-defined. But what about P(E|D)? Assuming the enzyme in question constitutes an unprecedented biological innovation, how do we assign a probability to a designer designing it?

The difficulty here is not confined to biological design hypotheses. Indeed, it applies to all cases of innovative design. To be sure, there are design hypotheses that confer reliable probabilities. For instance, my typing this book confers a probability of about thirteen percent on the letter "e"—that's how often on average writers in English employ the letter "e." But what's the probability of me writing this book? What's the probability of Rachmaninoff composing his variations on a theme of Paganini? What's the probability of Shakespeare writing his sonnets? When the issue is creative innovation, the very act of expressing the likelihood P(E|D) becomes highly problematic and prejudicial. It puts creative innovation by a designer in the same boat as natural laws, requiring of design a predictability that's circumscribable in terms of probabilities. But designers are inventors of unprecedented novelty, and such creative innovation transcends all probabilities.

(3) The illusion of mathematical rigor. As I noted in the previous point, if E denotes the occurrence of a certain gene coding for a certain novel enzyme, then P(E|H) can reasonably be regarded as having a well-defined probability. Provided that the problem of assessing this probability is not too technically difficult, we may be able to evaluate it precisely or at least estimate an upper bound for it. But what about P(E|D)? What about probabilities like this more generally where a design hypothesis confers a probability on a creative innovation? Not only is there no reason to think that such probabilities make sense (see the previous point), but when

Bayesians reason with such probabilities, they do so without attaching any precise numbers to them. The probability P(E|D) functions as a placeholder for ignorance, lending an air or mathematical rigor to what really is just a subjective assessment of how plausible a design hypothesis seems to the person offering a Bayesian analysis.

(4) Eliminating chance without comparison. Within the Bayesian approach, statistical evidence is inherently comparative—there's no evidence for or against a hypothesis as such but only better or worse evidence for one hypothesis in relation to another. But that all statistical reasoning should be comparative in this way cannot be right. There exist cases where one and only one statistical hypothesis is relevant and needs to be assessed. Consider, for instance a fair coin (i.e., a perfectly symmetrical rigid disk with distinguishable sides) that you yourself are tossing. If you witness a thousand heads in a row (an overwhelmingly improbable event), you'll be inclined to reject the only relevant chance hypothesis, namely, that the coin tosses are independent and identically distributed with uniform probability.

Does it matter to your rejection of this chance hypothesis whether you've formulated an alternative hypothesis? I submit it does not. To see this, ask yourself when do you start looking for alternative hypotheses in such scenarios. The answer is, Precisely when a wildly improbable event like a thousand heads in a row occurs. So, it's not that you started out comparing two hypotheses, but rather that you started out with a single hypothesis, which, when it became problematic on account of a wild improbability (itself suggesting that Fisherian significance testing lurks here in the background), you then tacitly rejected it by inventing an alternative hypothesis. The alternative hypothesis in such scenarios is entirely ex post facto. It is invented merely to keep alive the Bayesian fiction that all statistical reasoning must be comparative.

(5) Backpedaling priors. As a variant of the last point, return to the earlier example of an urn with a million balls, one black and the rest white. As before, imagine that a fair coin is to be tossed if a white ball is randomly sampled from the urn but that a biased coin with probability .9 of landing heads is to be tossed otherwise. This time, however, imagine that the coin is tossed not ten times but ten thousand times and that each time it lands heads. The probability of getting ten thousand heads in a row with the fair coin is approximately 1 in 10^{3010} and with the biased coin approximately 1 in 10^{458} (with ten thousand tosses, heads are bound to turn up for either coin). A Bayesian analysis then shows that the probability that a white ball was selected is approximately 1 in 10^{2546} and the probability that the lone black ball was selected is 1 minus that minuscule probability.

Should we therefore, as good Bayesians, conclude that the black ball was indeed selected and that the biased coin was indeed flipped (the selection of the black ball being vastly more probable, given ten thousand heads in a row, than the selection of a white ball)? Clearly this is absurd. The probability of getting ten thousand heads in a row with either coin is vastly improbable, and it doesn't matter which urn was selected. The only sensible conclusion is that *neither* coin was randomly tossed ten thousand times. A Bayesian may therefore want to change the prior probability to introduce some doubt about whether the urn and subsequently one of the two coins were random sampled. But as in the previous point, we need to ask what induces us to change or reevaluate our prior probabilities. Not strictly Bayesian considerations but rather considerations of small probability based on chance hypotheses that, as first posed, admit no

alternatives. The alternatives need then to be introduced subsequently because Fisherian, not Bayesian, considerations prompt them.

(6) Independent empirical evidence for design. Bayesian theorists are often wedded to a Humean inductive framework in which design hypotheses require independent empirical evidence of a designer actually at work (i.e., the camera is running and the designer is—or at least in principle could be—caught on video tape) before design may be legitimately attributed. We saw in the last chapter that this restriction is not just artificial but in fact incoherent because induction cannot be the basis for identifying design, there being no way to get that induction up and running. Nevertheless, for Bayesians wedded to Hume, it is convenient to block a Bayesian analysis that might implicate design from even getting started by denying that certain design hypotheses—like a design hypothesis that appeals to an unevolved intelligence to explain biological complexity—could even in principle admit independent empirical evidence.

Thus, rather than face the problem of assessing prior probabilities in such cases, Bayesians wedded to Hume merely impose an additional restriction on the Bayesian framework stipulating, in effect, that the Bayesian framework may not be used for design hypotheses without independent empirical evidence of a designer. Strictly speaking, this restriction has no place within the Bayesian probabilistic apparatus (Bayes's theorem works regardless where the probabilities associated with a design hypothesis come from—just plug in the numbers), but it is now increasingly being invoked against intelligent design. For instance, whereas Elliott Sober in his 1993 edition of *Philosophy of Biology* (and thus before intelligent design had intellectual currency) allowed considerable freedom for Bayesian design inferences in biology, in the 2000 edition of that book (after intelligent design had created considerable waves) he closed off any design inference to a designer lacking independent empirical evidence. Thus, whereas the 1993 edition gave intelligent design a lease on life, the 2000 edition took it away.

The independent empirical evidence requirement raises a curious dilemma for Darwinism. Imagine space travelers show up loaded with unbelievably advanced technology. They tell us (in English) that they've had this technology for hundreds of millions of years and give us solid evidence of it (perhaps by pointing to some star cluster hundreds of millions of light years away whose arrangement signifies a message that confirms the aliens' claim). Moreover, they demonstrate to us that with this technology they can atom by atom and molecule by molecule assemble the most complex organisms. Suppose we have good reason to think that these aliens were here at key moments in life's history (e.g., at the origin of life, the origin of eukaryotes, the origin of metazoans, and the origin of the animal phyla in the Cambrian). Suppose further that in forming life from scratch the aliens would not leave any trace (their technology is so advanced that they clean up after themselves perfectly—no garbage or any other signs of activity would be left behind). Suppose, finally, that none of the facts of biology are different from what they are now. Should we think that life at key moments in its history was designed?

We now have all the independent empirical evidence we could want for the existence of physically embodied designers capable of bringing about the complexity of life on earth. If in addition our best probabilistic analysis of the biological systems in question tells us that unguided natural processes could not have produced them with anything like a reasonable probability, is a Bayesian design inference now warranted? Could the design of life in that case

become more probable than a Darwinian explanation (probabilities here being interpreted in a Bayesian or likelihood sense) simply because independent empirical evidence attests to designers with the capacity to produce biological systems?

This prospect, however, should raise a worry for Darwinists. The facts of biology, after all, have not changed. Yet design would be a better explanation if designers capable of, say, producing the animal phyla of the Cambrian could be attested through independent empirical evidence. Note that there's no smoking gun here (no direct evidence of alien involvement in the fossil record, for instance). All we know by observation is that beings with the power to generate life exist and could have acted. Would it help to know that the aliens really like building carbon-based life? But how could we know that? Do we simply take their word for it? The data of biology and natural history, we assume, stay as they are now.

But if design is a better explanation simply because of independent empirical evidence of technologically advanced space aliens, why should it not be a better explanation absent such evidence? If Darwinism is so poor an explanation that it would cave the instant space aliens capable of generating living forms in all their complexity could be independently attested, then why should it cease to be a poor explanation absent those space aliens? Again, the facts of biology themselves have not changed.

Is there a way to salvage the independent empirical evidence requirement? Clearly it would be illegitimate to modify this requirement by ruling out circumstantial evidence entirely and permitting only direct "eye-witness" evidence of a designer actually manipulating the designed object in question. Even Elliott Sober would not go along with this proposal (see his *Reconstructing the Past*—to reconstruct the past we need circumstantial evidence). For Sober, circumstantial evidence could in principle support a biological design hypothesis. The important thing for Sober is that there be independent empirical evidence for the existence of a designer. But no smoking gun is required. In fact, to require a smoking gun in the sense of direct "eye-witness" evidence would be just as bad for Darwinism as for intelligent design. The evidence is just as circumstantial for one as for the other.

But once the independent empirical evidence for design can be circumstantial, establishing merely the existence of a designer with the causal power and opportunity to produce the effect in question (as in the alien thought experiment), we have exactly the same set of data to explain that we did before we acquired that evidence. The requirement for independent empirical evidence is therefore either vacuous (if it can be circumstantial) or prejudicial (if required to be direct). And in either case it obstructs inquiry into any actual design that might be present. If we require independent empirical evidence of design but don't have it, we won't see design even if it is there.

(7) Implicit use of specifications. And finally we come to the most damning problem facing the Bayesian approach, namely, that it presupposes the very account of specification and rejection region that it was meant to preclude. Bayesian theorists see specification as an incongruous and dispensable feature of design inferences. For instance, Timothy and Lydia McGrew regard specification as having no "epistemic relevance" (Symposium on Design Reasoning, Calvin College, May 2001). At that same symposium Robin Collins, also a Bayesian, remarked: "We

could roughly define a specification as any type of pattern for which we have some reasons to expect an intelligent agent to produce it." Thus a Bayesian use of specification might look as follows: given some event E and a design hypothesis D, a specification would assist in inferring design for E if the probability of E conditional on D is increased by noting that E conforms to the specification (which, á la Collins, is a "pattern for which we have some reasons to expect an intelligent agent to produce it").

But there's a crucial difficulty here that Bayesians invariably sidestep. Consider the case of the New Jersey election commissioner Nicholas Caputo accused of rigging ballot lines. (This example appears in a number of my writings and has been widely discussed on the Internet. A ballot line is the order of candidates listed on a ballot. It is to the advantage of a candidate to be listed first on a ballot line because voters tend to vote more readily for such candidates.) Call Caputo's ballot line selections the event E. E consists of 41 selections of Democrats and Republicans in sequence with Democrats outnumbering Republicans 40 to 1. For definiteness, let's assume that Caputo's ballot line selections looked as follows (newspapers covering the story to my knowledge never reported the actual sequence):

DDDDDDDDDDDDDDDDDDDDDRDDDDDDDDDDDDDDDD

Thus we suppose that for the initial 22 times, Caputo chose the Democrats to head the ballot line; then at the 23rd time, he chose the Republicans; after which, for the remaining times, he chose the Democrats.

If Democrats and Republicans were equally likely to have come up (as Caputo claimed), this event has probability approximately 1 in 2 trillion. Improbable, yes, but by itself not enough to implicate Caputo in cheating. Highly improbable events after all happen by chance all the time—indeed, any sequence of forty-one Democrats and Republicans whatsoever would be just as unlikely. What, then, additionally do we need to confirm cheating (and thereby design)? To implicate Caputo in cheating it's not enough merely to note a preponderance of Democrats over Republicans in some sequence of ballot line selections. Rather, one must also note that a preponderance as extreme as this is highly unlikely. In other words, it wasn't the event E (Caputo's actual ballot line selections) whose improbability the Bayesian needed to compute but the composite event E* consisting of all possible ballot line selections that exhibit at least as many Democrats as Caputo selected. This event—E*—consists of 42 possible ballot line selections and has improbability 1 in 50 billion. It's this event and this improbability on which the New Jersey Supreme Court rightly focused when it deliberated whether Caputo had in fact cheated. Moreover, it's this event that the Bayesian needs to identify and whose probability the Bayesian needs to compute to perform a Bayesian analysis.

But how does the Bayesian identify this event? Let's be clear that observation never hands us composite events like E* but only elementary outcomes like E (i.e., Caputo's actual ballot line selection and not the ensemble of ballot line selections as extreme as Caputo's). But whence this composite event? Within the Fisherian framework the answer is clear: E* is the rejection region (and therefore specification) that counts the number of Democrats selected in 41 tries. That's what the court used and that's what Bayesians use. Bayesians, however, offer no account of how they identify the events to which they assign probabilities. If the only events they ever considered were elementary outcomes, there would be no problem. But that's not the case.

Bayesians routinely consider such composite events. In the case of Bayesian design inferences (and Bayesians definitely want to draw a design inference with regard to Caputo's ballot line selections), those composite events are given by specifications.

Let me paint the picture more starkly. Consider an elementary outcome E. Suppose initially we see no pattern that gives us reason to expect an intelligent agent produced it. But then, rummaging through our background knowledge, we suddenly see a pattern that signifies design in E. Under a Bayesian analysis, the probability of E given the design hypothesis suddenly jumps way up. That, however, isn't enough to allow us to infer design. As is usual in the Bayesian scheme, we need to compare a probability conditional on design to one conditional on chance. But for which event do we compute these probabilities? As it turns out, not for the elementary outcome E, but for the composite event E* consisting of all elementary outcomes that exhibit the pattern signifying design. Indeed, it does no good to argue for E being the result of design on the basis of some pattern unless the entire collection of elementary outcomes that exhibit that pattern is itself improbable on the chance hypothesis. The Bayesian therefore needs to compare the probability of E* conditional on the design hypothesis with the probability of E* conditional on the chance hypothesis.

The bottom line is this: The Bayesian approach to statistical rationality is parasitic on the Fisherian approach and can properly adjudicate only among hypotheses that the Fisherian approach has thus far failed to eliminate. In particular, the Bayesian approach offers no account of how it arrives at the events upon which it performs a Bayesian analysis. The selection of those events is highly intentional, and in the case of Bayesian design inferences needs to presuppose an account of specification. Specified complexity, far from being refuted by the Bayesian approach, is therefore implicit throughout Bayesian design inferences.