NOÛS

NOÛS 47:1 (2013) 154–180

Multiple Studies and Evidential Defeat

MATTHEW KOTZEN
University of North Carolina at Chapel Hill

1 The Puzzle

You read a study in a reputable medical journal in which the researchers report a statistically significant correlation between peanut butter consumption and low cholesterol. Under ordinary circumstances, most of us would take this study to be at least some evidence that there is a real causal connection of some sort (hereafter: a **real connection**) between peanut butter consumption and low cholesterol.

Later, you discover that this particular study isn't the only study that the researchers conducted investigating a possible real connection between peanut butter consumption and some health characteristic. In fact, their research is being funded by the Peanut Growers of America, whose explicit goal is to find a statistical correlation between peanut butter consumption and some beneficial health characteristic or other that they can highlight in their advertising. Though the researchers would never falsify data or do anything straightforwardly scientifically unethical,[1] they have conducted one thousand studies attempting to find statistically significant correlations between peanut butter consumption and one thousand different beneficial health characteristics including low incidence of heart disease, stroke, skin cancer, broken bones, acne, gingivitis, chronic fatigue syndrome, carpal tunnel syndrome, poor self-esteem, and so on.

When you find out about the existence of these other studies, should that at least partially undermine your confidence that there is a real connection between peanut butter consumption and low cholesterol? In other words, does the existence of the other studies at least partially *defeat* the evidence that the study provides for a real connection between peanut butter consumption and low cholesterol? That is the question that this paper will address. Call a defender of an answer of "yes" to this kind of question a **Defeatist**, and call a defender of the opposite view an **anti-Defeatist**.

A few clarifications: First, there is the familiar point that A correlating with B isn't the same as A causing B, so it isn't obvious that the original study ever gives us any reason to think that eating peanut butter *causes* people to have low cholesterol; perhaps having low cholesterol causes people to eat peanut butter, or perhaps some other characteristic (like being tall) causes people both to eat peanut butter and to have low cholesterol.[2] So let's just focus on the question of whether there is *some real causal connection or other*[3] between peanut butter consumption and low cholesterol. It's very plausible that the study is evidence that there is such a real connection, whatever its nature might be.

Second, a Defeatist needn't think that the existence of the other studies *completely* defeats the evidence that the original study provides for a real connection between peanut butter consumption and low cholesterol; perhaps she thinks that she should be only *somewhat* less confident (and perhaps even still quite confident) in that connection after finding out about the other studies. This view still counts as a Defeatist one; to be an anti-Defeatist, you have to think that finding out about the existence of the other studies should do *nothing at all* to defeat your confidence that there is a real connection between peanut butter and low cholesterol.

Third, in some cases, there is reason to think that it might matter whether we have access to all of the details of the study which found a statistically significant correlation between peanut butter and low cholesterol (i.e. to all of the data on which the relevant statistical tests were based), or whether this data has been "black-boxed" in such a way that we have access only to the fact that the data were statistically significant (but *not* to the data themselves).[4] For instance, suppose that a judge hears some evidence at a criminal trial and on that basis judges that the defendant is guilty, and suppose that we too form the judgment that the defendant is guilty. Now, suppose that we learn that the judge always convicts people, regardless of the evidence against them. Is this a defeater of our evidence that the defendant is guilty? Here, it seems to matter whether, in forming our judgment that the defendant was guilty, we had access to all of the incriminating evidence, or whether we had access only to the fact that the judge found the defendant guilty. If we didn't have access to the incriminating evidence ourselves but rather formed our judgment that the defendant was guilty solely on the basis of the judge's finding, then it's fairly clear that the information that the judge always convicts should make us less confident that the defendant is guilty. By contrast, if we had access to all of the incriminating evidence and formed our own judgment on the basis of that evidence, then it is much less clear that the information that the judge always convicts is a defeater; after all, the reliability of the judge wasn't an essential premise in our reasoning in this case. So, to sidestep this sort of complication, let's assume that, in reading the journal study reporting a statistically significant correlation between peanut butter consumption and low cholesterol, we also came to learn all of the data on which the statistical test was based. The Defeatist view is that, even if we've seen all of the data, we should still become less confident in the existence of a real connection between peanut butter and low cholesterol once we learn that that data was collected in the manner described above.

Apart from its relevance to the interpretation of medical and other scientific studies, our puzzle is intimately related to several other puzzles in epistemology, metaphysics, and the philosophy of science. For instance, just as we will investigate whether the existence of many studies defeats the evidence that any one particular study provides for its conclusion, other philosophers have investigated whether the existence of many universes would defeat the evidence that the "fine-tuning" of our universe for life may provide for the hypothesis that our universe was designed by some intelligent agent.[5] Also, one might think that the fact that the researchers conducted many studies in an attempt to find some correlation or other is good

reason to think that the hypothesis of a real connection between peanut butter and low cholesterol *accommodated*, rather than *predicted*, the data, and it is a controversial question in the philosophy of science whether predicted data does more to confirm a theory than accommodated data.[6] I can't hope to settle all of these questions here. What I do plan to do is to argue for a solution to one of these related puzzles and hope that it takes us some of the way toward seeing a solution to the others.

I have three main goals in this paper. First, I will defend anti-Defeatism in a large class of cases about which many people have Defeatist intuitions. Second, I will spell out the very limited circumstances under which Defeatism is true. And third, I will explain away the Defeatist intuitions that many people (myself included) have even in cases where Defeatism is false.

2 Some Motivations for Defeatism

First, some terminology. Call the (alleged) real connection between peanut butter consumption and low cholesterol a **PB-LC Connection**, and call the study investigating that (alleged) real connection the **PB-LC Study**. Call the 999 other studies investigating possible real connections between peanut butter consumption and other health characteristics the **999 Other Studies**. Call the scenario in which only the PB-LC Study is conducted the **Single Study Scenario**. Call the scenario in which all thousand studies (i.e., the PB-LC Study and the 999 Other Studies) are conducted the **Multiple Study Scenario**. Call the (perhaps defeated) evidence provided by the PB-LC Study for the existence of a PB-LC Connection the **PB-LC Evidence**. Defeatism, then, is the view that finding out that you're in the Multiple Study Scenario rather than the Single Study Scenario is at least a partial defeater for the PB-LC Evidence.

Most people have some sort of Defeatist intuition about the peanut butter case. After all, it's very natural to think that in the Multiple Study Scenario, the researchers were bound to find *some* statistically significant correlation or other, regardless of whether there are any real connections between peanut butter and any of the health effects studied. Suppose that the PB-LC Study yielded data that statisticians would characterize as "significant at the $\alpha = .01$ level"; this is supposed to correspond to a 1% chance of collecting the relevant data[7] by chance if there is no real connection between peanut butter and low cholesterol.[8] Thus, if the researchers conducted one thousand studies investigating possible real connections between a completely inert sugar pill and various health characteristics, they would collect data statistically significant at the $\alpha = .01$ level in approximately 10 (i.e., 1%) of those studies. Since there is no independent reason to think that a real connection between peanut butter consumption and low cholesterol is especially likely, or that the study investigating that *particular* connection is especially reliable, it's very natural to think that learning of the existence of the 999 Other Studies at least partially defeats the PB-LC Evidence.

Classical statisticians seem, in general, to endorse this line of thought. Here is a representative passage from McCabe and Moore's *Introduction to the Practice of Statistics*:

> We will state it as a law that any large set of data—even several pages of a table of random digits—contains some unusual pattern. Sufficient [numbers of statistical tests] will discover that pattern, and when you test specifically for the pattern that turned up, the result will be significant. It also will mean exactly nothing.... It is convincing to hypothesize that an effect or pattern will be present, design a study to look for it, and find it at a low significance level. It is not convincing to search for any effect or pattern whatever and find one.[9,10]

McCabe and Moore go on to describe a famous case where a team of psychiatrists studied a group of schizophrenic patients and another group of non-schizophrenic patients, measuring 77 different variables and testing each to see whether it correlated with schizophrenia. The psychiatrists found two variables for which they could establish significance at the $\alpha = .05$ level and published their results. McCabe and Moore's judgment of this procedure is that "[r]unning one test and reaching the $\alpha = .05$ level is reasonably good evidence that you have found something; running 77 tests and reaching that level only twice is not."[11] This judgment is orthodoxy among classical statisticians.[12]

Moreover, there is a well-established classical statistical methodology for taking the number of tests being conducted into consideration, known as the "Bonferroni Correction."[13] The Bonferroni Correction entails that data that would ordinarily be significant at one level is only significant at some higher (i.e., *less* significant) level when multiple tests are conducted on the same data. More specifically, the Bonferroni Correction entails that, when we conduct 20 tests looking for correlations between two variables, and one of the tests would ordinarily be significant at the .01 level, the *real* significance level (known as the "per family" or "experimentwise" significance level) of the data is $.01 \times 20 = .20$; in this context, if we want data that is *actually* significant at the .01 level, then we need to look for data that would *ordinarily* be significant at the much lower (i.e., *more* significant) $\frac{.01}{20} = .0005$ level.

To justify the Bonferroni Correction, statisticians typically appeal to considerations similar to the ones appealed to by McCabe and Moore above. Here is Hervé Abdi, in the *Encyclopedia of Measurement and Statistics*:

> The more tests we perform on a set of data, the more likely we are to reject the null hypothesis [i.e., the hypothesis that there is no real connection between the variables] when it is true.... This is a consequence of the logic of hypothesis testing: We reject the null hypothesis if we witness a *rare* event. But the larger the number of tests, the easier it is find rare events and therefore the easier it is to make the mistake of thinking that there is an effect when there is none. This problem is called the *inflation* of the [significance] level.... [T]he larger the number of experiments, the greater the probability of detecting a low-probability event.... In fact, waiting long enough is a sure way of detecting rare events!... In order to be protected from [this problem], one strategy is to correct the [significance] level when performing multiple tests.[14]

As applied to the peanut butter case, the thought here seems again to be that in the Multiple Study Scenario, even though it's unlikely that we'd find a statistically significant correlation specifically between peanut butter consumption and low cholesterol on the assumption that there is no real connection between them, it's quite likely that we'd find *some statistically significant correlation or other*, and that this second fact serves to at least partially defeat the PB-LC Evidence.

A second motivation for Defeatism is that it's quite natural to think that conducting multiple studies in an attempt to establish at least one statistically significant correlation is akin to a practice that looks to be even more obviously problematic: namely, using a biased "stopping rule" in a *single* study that tells the researcher not to stop collecting data until the desired result is obtained. For example, suppose that you read a study describing one group of 500 subjects who don't eat peanut butter and another group of 500 subjects who do eat peanut butter, and suppose that the study reports a statistically significant difference between the mean cholesterol levels of the two groups. Again, this is some evidence that there is a PB-LC Connection. But suppose that you later learn that the study was *not* originally designed to have 500 subjects in each group. Rather, the researchers originally constructed the two groups with 100 subjects each, but didn't find the statistically significant result that they were hoping for. So they added 100 subjects to each group, still found nothing, added 100 more, then 100 more, then 100 more, and finally found a statistically significant result with 500 subjects in each group, at which point they promptly ended the study.

Now, it is controversial[15] precisely how we ought to react epistemically to learning about the "biased" stopping rule in a case like the one above, and I don't intend to enter explicitly into that debate here.[16] But one natural reason to think that a biased stopping rule is illegitimate in a single study is that it increases the probability that the researchers would get *some statistically significant result or other*. To the extent that one thinks that this makes the biased stopping rule scientifically or epistemologically problematic, one might think that similar considerations apply to the meta-stopping rule "Keep conducting studies looking for correlations between peanut butter and various health characteristics until you get a statistically significant result (or your funding runs out)," which only the Defeatist looks to have the resources to criticize.

Finally, there is a wide range of thought experiments about which Defeatism is quite intuitively tempting. I introduce you to my pet monkey George, who has just typed a one-act play of staggering subtlety and dramatic merit. It's natural to think that this is at least some evidence for the hypothesis that George is far more intelligent than your average monkey. But if you were to find out that I have *one trillion* pet monkeys that I force to spend the entire day typing, you would likely be far less inclined to believe that George is a talented playwright; after all, it is a familiar point[17] that if you give enough monkeys enough time in front of a typewriter, one of them is *bound* to type something interesting eventually. But this Defeatism about multitudes of monkeys seems to be motivated by precisely the same considerations that motivate Defeatism about multitudes of studies.

3 Some Motivations for anti-Defeatism

While I think that Defeatism has a good deal of intuitive plausibility, there are a number of prima facie difficulties that it faces. In this section, I will raise six such difficulties. In the end, I think that each of these difficulties is insuperable for the Defeatist in the vast majority of cases, though for reasons of space I won't be able to explore every response that the Defeatist might offer.

First of all, it's not clear that all of the motivations for Defeatism discussed in Section 2 are particularly strong. For instance, despite the *existence* and *wide use* among statisticians and research scientists of the Bonferroni Correction, we still might wonder about its justification. The intuitive justification seems just to be the informal one sketched in Section 2, but the considerations that motivate the correlation test procedure *itself* seem *not* to motivate the Bonferroni Correction; otherwise we wouldn't need the Bonferroni Correction as an additional "axiom" in the statistical framework, for the Correction would follow from the basic principles of that framework. If the statistical framework that underwrites correlation testing is legitimate, there's a real question about why the correlation test procedure that it justifies can be applied without correction only once. The mathematical procedure for taking the derivative of a function works as often as we like, as does the rule for calculating a conditional probability, etc. One might think that there is something highly suspect about the fact that we can use the classical statistical method for calculating the significance level of a correlation test only once with respect to a certain body of data, even if statisticians are able to come up with a Correction that formalizes that restriction. We will return to this point in Section 4.

Second, Defeatism faces a version of the so-called **Generality Problem**.[18] The Generality Problem is the problem of specifying in some non-arbitrary way precisely which class of tests, trials, or events is relevant when assessing a particular piece of evidence. In the peanut butter case, the Defeatist's thought was that even on the assumption that there is no real connection between peanut butter consumption and any of the thousand health effects being investigated, it was still very likely that at least one of the thousand studies would yield a significance level of .01. But why should we stop there? Why should we consider only the thousand studies run by these particular researchers investigating peanut butter connections? After all, it's also quite likely that, given all the studies that these researchers have run in their *careers* investigating all sorts of different questions, at least one of *those tests* was going to yield a significance level of .01. Indeed, given all the studies that have been performed by members of the scientific community in the last hundred years, it's even more likely that at least one of those studies would yield a significance level of .01. And why fetishize the .01 level? Given all the studies that have been performed in the last hundred years, it's *overwhelmingly* likely that at least one of them would yield a significance level of .20 or lower. In other words, the PB-LC Study is a token of the peanut-butter-cholesterol-study type, the peanut-butter-health-characteristic-study type, the study-run-by-these-researchers type, the study-run-in-the-past-100-years type, the study-with-a-.01-significance-level type, the study-with-a-.20-or-lower-significance-level type, the study-run-on-a-Tuesday-in-June-type (say), etc. The Generality Problem is the problem of specifying

precisely which of these types is relevant when we consider the probability that some token of that type would produce the result in question.[19]

Third, and relatedly, there is what I'll call the **Triviality Problem**. The Defeatist seems to be endorsing the strategy of **weakening the evidence**. She notices that the PB-LC Study yields a significance value of .01. She then goes on to notice that though this *particular* event was very unlikely to occur by chance in the Multiple Study Scenario, it was much more likely that *at least one of the thousand studies would yield a significance value of .01*. Thus, she seems to be weakening the evidence from **The PB-LC Study yielded a significance value of .01** to the weaker **At least one of the thousand correlation tests yielded a significance value of .01**.[20]

The Triviality Problem is that this move of weakening the evidence seems to be too powerful, allowing us to dismiss almost *any* event as a coincidence, regardless of its likelihood on the chance (null) hypothesis.[21] We notice that blood matching the suspect's DNA was found at the murder scene, and at first take this to be evidence that the suspect was at the murder scene. After all, it's very unlikely that this suspect's DNA would end up at this murder scene by chance if he wasn't there. But suppose that we weaken the evidence from **This particular suspect's DNA was found at this particular murder scene** to **Somebody's DNA was found at some murder scene at some point or other in the last hundred years**. Though the former, stronger piece of evidence is quite unlikely to be collected by chance, the latter, weaker piece of evidence is quite likely indeed to be collected by chance. Still, it's not clear that this fact does *anything at all* to exonerate the suspect in this particular case. Something has clearly gone wrong with the strategy of weakening the evidence. The Triviality Problem is to say what.

Fourth, there is what I'll call the **Independence Problem**. The results of each of the thousand studies conducted in the peanut butter case are naturally taken to be *independent* of each other; the results of any one study don't give us any information about how the other studies will turn out.[22] Moreover, the results of any particular study are independent of the *existence* of the other studies; regardless of whether the researchers also decided to conduct the 999 Other Studies, the probability that they'd find a statistically significant result in the PB-LC Study is the same.

A number of cases illustrating the consequences of this sort of independence have been discussed in the fine-tuning literature mentioned in Section 1, but I think that they are instructive here as well. Suppose that you know that either only your friend Joe is going to roll a pair of dice once in his office, or that Joe and 999 other people are each going to roll a pair of dice once in their respective offices. Learning that all one thousand people rolled dice is obviously evidence that someone rolled double-sixes (since there were one thousand opportunities to roll double-sixes rather than just one). However, equally obviously, learning that all one thousand people rolled dice is *not* evidence that *Joe* rolled double-sixes. *Joe's* chances of rolling double-sixes were the same (i.e., $\frac{1}{36}$) regardless of whether other people were rolling dice in their offices as well (and regardless of how those dice landed if they were rolled).[23]

The analogy to the peanut butter case should be clear. Just as the information that all one thousand people rolled dice is evidence that *someone* rolled double-sixes,

so too is the fact that one thousand studies were conducted evidence that *at least one of them* would yield a significance level of .01. But just as the information that all one thousand people rolled dice is no evidence at all that *Joe* would roll double-sixes, neither is the fact that one thousand studies were conducted any evidence at all that *the PB-LC Study* would yield a significance level of .01; regardless of whether the other studies are conducted or not, the probability that the PB-LC Study would yield a significance level of .01 is the same. As a result, there's no obvious way in which the existence of the 999 Other Studies could possibly be relevant to the evidential impact of the PB-LC Study, as the Defeatist claims.

Fifth, it is at least somewhat natural to think that, as long as we're sure that the data hasn't been falsified or altered in some other uncontroversially misleading manner, an evaluation of the evidential impact of a study shouldn't require an investigation into the *goals* or *intentions* or *hopes* or *histories* of the researchers who conducted that study. If we find a researcher's laboratory notes from a particular study but can't locate her for some reason, Defeatism seems to entail that we should ask her friends, family, and colleagues whether she ran other studies besides that one or not; if she ran others, Defeatism entails that the study is less significant than if she didn't. Moreover, Defeatism looks to be committed to the evidential relevance even of *counterfactual* scenarios. Suppose again that we find a researcher's laboratory notebook from a study but can't locate her, and this time suppose that we know that she ran only one study. Still, it seems as though Defeatism is committed to the relevance of her *plans* or *intentions*—would she have kept conducting more and more studies if she didn't like the results of the one that she in fact ran? After all, if she would have done so,[24] then she was bound to find some statistically significant result or other (even if she ended up *actually* conducting only one study), and the same reasoning that motivates Defeatism about multiple studies seems to similarly motivate Defeatism about plans or intentions of this sort. But this conclusion is rather odd; we don't ordinarily think that the private psychological states of researchers are relevant in this way. Call this the **Psychology Problem**.

Sixth, and finally, there is a worry that Defeatism entails an unacceptable sort of non-commutativity of evidence. Here's why: Suppose that, at 1:00 pm, we observe the PB-LC Study being performed, and we see that it yields a statistically significant result. The Defeatist and the anti-Defeatist agree that that is evidence for the existence of a PB-LC Connection. Next, suppose that at 2:00 pm, we observe the 999 Other Studies being performed. Should the mere fact that the 999 Other Studies were performed *after* the PB-LC Study yielded a statistically significant result be a reason to lower our credence in the existence of a real PB-LC Connection? That would be surprising. For reasons related to the fifth motivation for anti-Defeatism above, we ordinarily think that it's perfectly acceptable for scientists to publish studies reporting statistically significant correlations, even if they plan to *later* conduct other studies investigating possible connections involving one of the variables in the study they're publishing. If a scientist conducts one study and finds a statistically significant correlation between some drug D and (say) flu symptoms, and publishes the results, we don't think that he should then publish a *retraction* if he goes on to perform other studies investigating possible connections

between D and symptoms of other diseases (regardless of the results of those latter studies). For instance, we have all the evidence we need right now that smoking is really connected with lung cancer, and this evidence wouldn't be diminished in the slightest if researchers decided tomorrow to conduct billions of studies investigating possible connections between smoking and other diseases.

But, if we accept the claim that the existence of the 999 Other Studies doesn't do anything to defeat the PB-LC Evidence in the scenario where the 999 Other Studies are performed *after* the PB-LC Study, why should we think that the existence of the 999 Other Studies has any stronger of a defeating effect if those studies are performed *before*, or *at the same time as*, the PB-LC Study?

Most epistemologists[25] accept some version of the

> **Principle of the Commutativity of Evidence** (COMMUTE): Facts about the order in which one's evidence is acquired shouldn't make any difference to what it is reasonable for her to believe.

There are some complications in formulating COMMUTE that are beyond the scope of this paper to address,[26] but there seems to be a consensus among epistemologists that these complications can be sidestepped.[27] And COMMUTE is quite intuitive in the vast majority of cases; if I'm trying to decide how confident I should be that it will rain tomorrow, for example, it shouldn't matter whether I hear the weather report on the news first and then read the newspaper forecast second, or vice versa.

If COMMUTE is correct as applied to this situation, then anti-Defeatism seems to follow. For if the PB-LC Study is evidence for the existence of a PB-LC Connection in the Single Study Scenario, and if the researcher's performing the 999 Other Studies *after* the PB-LC Study doesn't do anything to defeat that evidence, and if (as COMMUTE entails) our epistemic situation with respect to the claim that there is a PB-LC Connection doesn't change in the scenario where the very same studies are performed in a different order, then we should have just as much evidence for a PB-LC Connection in the scenario in which the PB-LC Study is the 139th or the 726th study conducted as we do in the scenario in which it is the 1st. And if *that's* right, then finding out that the 999 Other Studies were performed in addition to the PB-LC Study shouldn't do anything to defeat the PB-LC Evidence. In other words, Defeatism must be false. Call this the **Commutativity Problem**.

4 A Toy Case

Suppose that we're not yet convinced by the arguments for anti-Defeatism presented in Section 3. How else might we try to resolve the dispute between the Defeatist and the anti-Defeatist? One thing we might try is to think carefully through a more precisely specified toy case that has a number of features in common with the peanut butter case, and see whether Defeatism can be sustained in that case.[28]

Thus: Consider a jar containing 1,000,000 well-mixed dice, 99% (i.e., 990,000) of which are fair and 1% (i.e., 10,000) of which are biased in such a way that they always land 6. Now, suppose that a die (call it "Harry") is collected at random and

rolled three times in a row, landing 6 all three times. How confident should we be that Harry is biased?

If we're Defeatists, then we'll think that the case is crucially under-described. For a Defeatist, knowing merely that Harry was collected at random *in some sense of the term "random"* isn't enough to know how we should set our credence that Harry is biased. If Harry was the *only* die that was selected from the jar and rolled, then perhaps we would have good reason to believe that Harry is one of the biased dice (after all, the probability that any particular fair die will land 6 three times in a row is $(\frac{1}{6})^3 = \frac{1}{216} \approx 0.00463$, whereas a biased die that is tossed three times is *certain* to land 6 all three times). But if some large number of dice (say, 1,000, including Harry) were all randomly collected from the jar at once,[29] and they were all tossed three times, and all we know about Harry is that it was *one* of the dice that landed 6 all three times, then we would have less reason to believe that Harry is biased. After all, given that we collected 1,000 dice and rolled them *each* three times, the probability is quite high (approximately[30] .99) that *at least one* of the collected dice would land 6 three times in a row, even if every single one of them is in fact fair (and, the Defeatist might go on to point out, there's no *special* reason, beyond the fact that it landed 6 three times, to think that *Harry* is biased). Thus, if we take his reasoning in the peanut butter case as a fair guide, the Defeatist would claim that we should be less confident that Harry is one of the biased dice if 1,000 dice (including Harry) were drawn from the jar and rolled than if just Harry was.

An anti-Defeatist, on the other hand, doesn't think that the case is under-described in the way that the Defeatist takes it to be. An anti-Defeatist's credence that Harry is biased won't depend on whether Harry was drawn from the jar by itself or along with 999 other dice, since the anti-Defeatist is straightforwardly committed to the irrelevance of this information about the number of "studies." All that matters, for the anti-Defeatist, is the fact that Harry was rolled three times and that it landed 6 all three times (together with her knowledge of the composition of the jar and of the fact that Harry was drawn at random from that jar).

Call the claim that Harry is one of the biased dice **HarryIsBiased** and call the claim that Harry has been rolled three times in a row and landed 6 all three times **Three6s**. Call the scenario in which Harry was drawn from the jar alone the **Single Drawing Scenario** and call the scenario in which Harry was drawn from the jar along with 999 other dice the **Multiple Drawing Scenario**. Then, the Defeatist is committed to being less confident in **HarryIsBiased** after learning **Three6s** in the Multiple Drawing Scenario than she is in the Single Drawing Scenario. There are two ways that the Defeatist might elaborate such a sensitivity to the Drawing Scenario.

First, depending on the Drawing Scenario, she could assign a different *prior* probability to **HarryIsBiased**—assigning a *lower* prior probability to **HarryIsBiased** in the Single Drawing Scenario than in the Multiple Drawing Scenario. That way, even though updating on **Three6s** will increase her confidence in **HarryIsBiased** regardless of how Harry was drawn from the jar, such a Defeatist will end up with a lower posterior credence in **HarryIsBiased** in the Multiple Drawing Scenario, since she had a lower prior credence in **HarryIsBiased** in that scenario.

But this approach is curious, to say the least. After all, we know that the jar contains a large number of (stipulatively randomized) dice, and we know that 99% of them are fair and that 1% of them are biased. Presumably, even the Defeatist wants to allow that when you pick *just one* die (let's suppose it's Harry) at random from a jar with such a composition of dice, a rational agent's prior credence in **HarryIsBiased** should be .01.[31] So, a Defeatist adopting this approach must want to claim that a rational agent's prior credence in **HarryIsBiased** should be *lower than .01* in the Multiple Drawing Scenario. But why should we accept that? If you collect 1,000 dice out of a jar containing dice, 1% of which are biased, isn't it obvious that your credence that any particular die is biased should be .01? This is certainly how we reason in other contexts. A doctor knows that 1% of the population has a particular genetic condition C, and she has no information beyond that about whether patient P has C. Should her *prior* credence that P has C (i.e., her credence *before* examining P or performing any tests at all on P) be affected by whether P was sitting in the doctor's waiting room alone or with 999 other people? And does it really seem rational for the doctor to have any credence other than .01 that P has C at this point?

Moreover, imagine that we collected *all 1,000,000 dice* from the jar all at once. We *know* that 990,000 of these dice are fair and that 10,000 of them are biased; that much is just stipulated in the construction of the example. Does the Defeatist really want to claim, for each die in the sample of 1,000,000, that the rational credence that it is biased is lower than .01? It's not clear that such a position is even coherent. After all, the Defeatist I'm imagining has already agreed that if we collect *just one* die at random from the jar, the rational credence that it is biased is .01. But now he wants to claim that, when we collect all 1,000,000 dice at the same time, our credence that any particular die is biased should be lower than .01. But we can always *re*-sample *just one die* from the collection of 1,000,000 dice that we've just collected from the jar; what should our credence that that die is biased be? The Defeatist I'm imagining seems committed both to the claim that it should be .01 and to the claim that it should be lower than .01. So I don't think this is the most promising way for the Defeatist to elaborate his view.

The second—and at least somewhat more promising—way for the Defeatist to go is to agree with the anti-Defeatist that the rational *prior* credence (before learning **Three6s**) to assign to **HarryIsBiased** is .01, regardless of whether Harry was drawn from the jar by itself or with 999 (or 999,999) other dice. Thus, a Defeatist adopting this approach avoids all of the problems discussed above with the first Defeatist approach, since his *prior* credence that any particular die is biased when it is drawn from the jar at random is .01, regardless of the Drawing Scenario.

But if this approach is going to qualify as a Defeatist one, it must recommend a lower *posterior* credence (after learning **Three6s**) in **HarryIsBiased** in the Multiple Drawing Scenario than it does in the Single Drawing Scenario. And since the Defeatist approach under consideration assigns the *same* prior to **HarryIsBiased** in each Drawing Scenario, this approach is committed to recommending a procedure for updating on new evidence (such as **Three6s**) that is somehow sensitive to the Drawing Scenario. That's the only way for it to be

possible for the update procedure to take as inputs the same priors for **HarryIsBiased** in the two Drawing Scenarios and yet to yield different posteriors as outputs. But it's not at all obvious what such an update procedure could be. It certainly isn't the Bayesian Rule of Conditionalization;[32] that rule is sensitive only to the value of the prior p(**HarryIsBiased**) and to the values of the likelihoods p(**Three6s** | **HarryIsBiased**) and p(**Three6s** | ¬ **HarryIsBiased**). The Defeatist under consideration has already conceded that the prior p(**HarryIsBiased**) is the same in each Scenario. The former likelihood p(**Three6s** | **HarryIsBiased**) is obviously 1 in either Drawing Scenario, since biased dice are guaranteed to land 6 on all three rolls (regardless of the Drawing Scenario), and the latter likelihood p(**Three6s** | ¬**HarryIsBiased**) is almost as obviously $(\frac{1}{6})^3$ in either Drawing Scenario, since a fair die will land 6 three times in a row with probability $(\frac{1}{6}) \times (\frac{1}{6}) \times (\frac{1}{6})$ (again, independently of the Drawing Scenario). Thus, the Bayesian Rule of Conditionalization just doesn't allow for any sensitivity at all to the Drawing Scenario. So this second Defeatist approach—however it is spelled out—will have to be manifestly non-Bayesian.

Of course, the *costs* of having a non-Bayesian update policy in general are somewhat controversial,[33] and I don't want to assume that *any* deviation from orthodox Bayesianism is a fault in the Defeatist view. But the Bayesian assumptions here are really quite modest,[34] and the Bayesian reasoning in this case is overwhelmingly plausible. Moreover, it can be shown that if you were to *bet* in accordance with the Defeatist policy, you should expect to lose money, whereas you should not expect to lose money betting in accordance with the anti-Defeatist policy.[35]

Also, it's not ultimately clear that this second Defeatist policy is any more coherent than the first. In the Single Drawing Scenario, the Defeatist assigns the same posterior credence as the anti-Defeatist to **HarryIsBiased**, whereas in the Multiple Drawing Scenario, the Defeatist assigns a lower posterior credence to **HarryIsBiased** than the anti-Defeatist. But if Defeatism is supposed to be *generally* true, then it seems as though the Defeatist policy has to be *generally applicable*; but the (perhaps somewhat obvious) problem is that the Single Drawing Scenario repeated again and again *just is* the Multiple Drawing Scenario. If the Defeatist thinks that the fact that several dice were drawn *at once* from the jar and rolled is a reason to be less confident that Harry is biased when it lands 6 three times in a row, he's also presumably going to think that the fact that many dice were drawn from the jar, *one at a time*, and rolled is similar reason for decreased confidence—but this difference is the *only* difference between the Single Drawing Scenario and the Multiple Drawing Scenario, at least when they're each repeated indefinitely often. Perhaps the Defeatist could come up with some grounds to object to his update policy in the Single Drawing Scenario being applied each time in such a repeated sequence. But then we are forced to wonder: if some conclusion is the right one to draw today, why shouldn't the exact same conclusion be the right one to draw tomorrow (assuming that no new relevant information has come to light)?

By contrast, the most reasonable anti-Defeatist policy in this context, it seems, is just the policy that a Bayesian agent would adopt—i.e., the policy of setting her

prior credences in accordance with the known composition of the jar and then updating on new information by Bayesian Conditionalization. Thus, given that the jar contains a large number of dice, 1% of which are biased so as to always land 6 and 99% of which are fair, her prior credence in **HarryIsBiased** is .01, regardless of whether Harry was drawn from the jar alone or along with 999 others. Then, when she learns **Three6s**, she updates her credence in **HarryIsBiased** from .01 to approximately[36] .69 by Conditionalization. Again, since nothing in the Bayesian calculation takes into consideration whether Harry was drawn from the jar alone or along with 999 other dice, the anti-Defeatist's final credence that Harry is biased will be insensitive to such considerations about the manner in which Harry was drawn from the jar. The anti-Defeatist therefore avoids all of the problems with Defeatism discussed above.

A Defeatist certainly *could* respond to the argument of this section by claiming that the peanut butter case with which we began is in some important sense disanalogous to the dice case, and that while perhaps the anti-Defeatist view is correct as applied to the dice case, the Defeatist view is correct as applied to more "realistic" cases such as the peanut butter case. Perhaps this is so—there certainly are *some* differences[37] between the peanut butter case and the dice case—but I don't see any reason to think that any of them is relevant. In both cases, we have reason to think that it's unlikely but not impossible that any particular [health characteristic/die] is really connected with [peanut butter/6s], we get some evidence for this connection from the [study/run of three 6s], and then we ask the question whether the plentitude of [studies/drawn dice] does anything to undermine this evidence. If there is an important disanalogy here, I think that the burden is on the Defeatist to say what it is and to explain why it makes the difference that it is alleged to make.

5 Selection Bias

In the peanut butter case, one potential reason for hesitation about the conclusion that there is a real PB-LC Connection might come from the thought that there's a *selection bias* involved in our access to the results of the studies. More specifically, one might worry that since the researchers look only for correlations between peanut butter consumption and *beneficial* health characteristics, and since they publish only those studies that yield statistically significant results, we have biased access to the results of the studies, much as we would have biased access to the average size of the fish in a pond if we sampled fish from the pond using a net that could catch only big fish.[38] It's natural to think that the discovery that our net catches only big fish would at least partially defeat the evidence that a big fish in the net provides for the hypothesis that the pond contains mostly big fish; the Defeatist might similarly argue that a selection bias present in the peanut butter case at least partially defeats the PB-LC Evidence.

The first and most important point to make about a selection bias in the peanut butter case is that it's an effect that is *independent* of the information that the 999 Other Studies were conducted. So if you have a general worry about the fact that only studies reporting statistically significant correlations are published in scientific

journals, that should be a worry that you have *even before* learning about the existence of the 999 Other Studies. In other words, even if it's correct that the fact that we read only studies reporting statistically significant results introduces a selection bias, we haven't yet seen any reason to think that this does anything to support the Defeatist view. After all, the fact that the 999 Other Studies were conducted doesn't obviously give us any *new* reason to think that the relevant selection bias is particularly strong, or operating in an especially biasing manner, or anything like that; the fact that some researchers have performed one thousand studies rather than just one doesn't make the editors of any scientific journal any less inclined to publish statistically insignificant results.

Second, there is a subtle but important difference between the selection bias introduced by a fish net that is too large to catch small fish and the selection bias present in the peanut butter case. In the fish net case, the selection bias guaranteed that we were going to collect some putative evidence in favor *of the hypothesis that the pond consists mostly of large fish*; since we were certain to catch a large fish regardless of whether the pond consists mostly of large fish or not,[39] our catching a large fish doesn't confirm the hypothesis that the pond consists mostly of large fish. This same effect *would* be present in the peanut butter case if the researchers had run one thousand studies looking specifically for a statistically significant correlation *between peanut butter and low cholesterol*. If they had done *that*, then it would have been overwhelmingly likely that at least one of the studies would have yielded a statistically significant correlation *between peanut butter and low cholesterol*, and the putative evidence that that correlation provided for a PB-LC Connection would have been defeated. But in the peanut butter case as presented, we're much more likely to observe a statistically significant correlation between peanut butter and low cholesterol if there is a real PB-LC Connection than if there isn't, since only *one* PB-LC study is run.

Third, none of my analysis of the dice case from Section 4 is affected by the introduction of a selection bias analogous to the one that prevents studies with statistically insignificant results from being published. In fact, there already was such a selection bias built into the case, since we considered the credence that the Defeatist and the anti-Defeatist would assign to the hypothesis that a given die is biased only in the case where that die landed 6 three times in a row. We could easily imagine that, if a die doesn't land 6 three times in a row, the agents *aren't even told about the roll*; again, precisely nothing changes about my analysis of each agent's credence that a die is biased *when it does land 6 all three times*.

So, I see no reason to think that the arguments presented above for anti-Defeatism are at all affected by the presence of a selection bias causing only certain results to be reported.

6 Independence

Up until now, I've been assuming that the results of the PB-LC Study are *independent* of both the existence and the results of the 999 Other Studies. Indeed, the Independence Problem introduced in Section 3 was precisely the problem of

how, given this independence, the 999 Other Studies could possibly be relevant to the PB-LC Evidence. How plausible is this independence assumption, and what happens if we relax it?

Well, the assumption that the results of the PB-LC Study are independent of the *existence* of the 999 Other Studies is quite plausible; finding out merely that the 999 Other Studies were conducted (without learning anything about the results of those studies) doesn't tell us anything at all about how the PB-LC Study is going to turn out. So I don't think that it's helpful to investigate the relaxation of this assumption. But it is more plausible that the results of the PB-LC Study might fail to be independent of the *results* of the 999 Other Studies; perhaps, for example, the fact that peanut butter consumption is statistically correlated with (say) low incidence of gingivitis is *some* reason to believe that peanut butter has generally healthful consequences, which is in turn *some* reason to think that peanut butter will also be statistically correlated with low cholesterol.

One still might worry, however, that this sort of non-independence wouldn't address the Independence Problem. After all, the Defeatist view is that finding out *merely about the existence* of the 999 Other Studies serves as a defeater for the PB-LC Evidence; as long as we still think that that the results of the PB-LC Study are independent of the existence of the 999 Other Studies, the Independence Problem seems to remain.

However, this worry fails to take account of the circumstances under which *we would learn* about the results of the 999 Other Studies. After all, as was implicit in Section 5 above, what we're actually learning when we find out about the results of the PB-LC Study isn't merely that **The PB-LC Study yielded a statistically significant result**, but rather that **We're learning that the PB-LC Study yielded a statistically significant result**.[40] It matters a great deal, then, whether we are in circumstances where these two pieces of information are equivalent or not. In particular, the relevant question is: what would happen if some of the 999 Other Studies yielded statistically significant results?

There are three possible answers here.

First, the researchers might have decided in advance to tell us only about a statistically significant result in the PB-LC Study (should one occur), ignoring any statistically significant results that arise in the 999 Other Studies. Call this **Situation 1**. In Situation 1, **The PB-LC Study yielded a statistically significant result** and **We're learning that the PB-LC Study yielded a statistically significant result** are equivalent, since we'd learn about a statistically significant result in the PB-LC Study if and only if one occurred. In this case, the information that the 999 Other Studies were conducted is truly irrelevant to the existence of a PB-LC Connection, even if the results of those 999 Other Studies are not independent of the results of the PB-LC Study. Suppose I get a run of three 6's on my die, and become somewhat more confident that my die is biased in favor of 6. Even if I'm certain that 999 other dice are biased (or unbiased) in exactly the same way as my die, still finding out that the 999 other dice were rolled (without finding out how they landed) gives me no information about their biasing, and hence gives me no new information about the biasing of my die. This is because, given this

setup, **My die had a run of 6's** and **I'm learning that my die had a run of 6's** are equivalent.

Second, the researchers might have decided in advance to tell us about all of the statistically significant results that arose in any of the thousand studies.[41] Call this **Situation 2**. In Situation 2, if the only statistically significant result that we learn about is the one in the PB-LC Study, then from the fact that we *haven't* been told about any statistically significant results in the 999 Other Studies, we can conclude that none of them was statistically significant, and hence we should become less confident that there are any real connections between peanut butter and any of the health effects investigated in the 999 Other Studies. And if we had antecedent reason to think that the results of the 999 Other Studies are non-independent *and positively correlated* with the results of the PB-LC Study,[42] then reason to think that there is no real PB-Gingivitis Connection (for instance) is also reason to think that there is no real PB-LC Connection (and same for the 998 other connections under consideration). Thus, in this case, finding out that the 999 Other Studies were conducted (and that we weren't told about any statistically significant results) is evidence against a PB-LC Connection, as the Defeatist claims.

Third, the researchers might have decided in advance to tell us about only some proper subset of the the statistically significant results that arose among the thousand studies. Call this **Situation 3**. For simplicity, let's focus on the case where the researchers decided in advance to tell us about *just one* statistically significant result, selected at random from among the statistically significant results of the thousand studies.[43] Now, suppose again that we have reason to believe that the existence of a PB-LC Connection and the existence of a PB-Gingivitis Connection are themselves positively correlated. Then we can see how, even though **The PB-LC Study yielded a statistically significant result** is independent of **The PB-Gingivitis Study was conducted**, it is *not* the case that **We're learning that the PB-LC Study yielded a statistically significant result** is independent of **The PB-Gingivitis Study was conducted**.

The reason is that, if there is a positive correlation between a PB-LC Connection and a PB-Gingivitis Connection, and if the PB-Gingivitis Study is conducted, then **The PB-LC Study yielded a statistically significant result** is some evidence for **The PB-Gingivitis Test yielded a statistically significant result**. Thus, **The PB-LC Study yielded a statistically significant result** is evidence that the PB-LC Study has additional "competition" from the PB-Gingivitis Study when the researchers randomly select *one* statistically significant test to tell us about; since the researchers select only one test to tell us about, we're less likely to learn about the statistically significant results of the PB-LC Study if the PB-Gingivitis Study yielded statistically significant results too. Thus, **We're learning that the PB-LC Study yielded a statistically significant result** is more likely to be true if the two Connections under consideration are independent than if they are positively correlated. As a result of this, **We're learning that the PB-LC Study yielded a statistically significant result** is better evidence for a PB-LC Connection if the Connections are independent than if they are positively correlated.[44] But this effect is not present if we know that the PB-Gingivitis Study isn't conducted (as in the Single Study Scenario).[45]

Thus, if there is reason to believe that the results of the PB-LC Study and the results of the 999 Other Studies are positively correlated, and if we're in either of Situation 2 or Situation 3, then the Defeatist is right.

Moreover, if we do have reason to believe that we are in either of Situation 2 or Situation 3, we are able to defend a Defeatist line while avoiding the six problems for Defeatism that I raised in Section 3.

First of all, we're not positing any kind of Correction as an axiom that prevents the Defeatist view from being generally applicable; the truth of Defeatism in the Multiple Study Scenario in either Situation 1 or Situation 2 is derived directly from the probabilistic tools that are used to analyze the Single Study Scenario.

The Generality Problem is solved too in either Situation 2 or 3. In both situations, the relevant class of studies to consider is just those studies that we might have found out about the results of, and which we have reason to believe are non-independent. So, other studies investigating correlations between peanut butter and health characteristics are relevant only if we have reason to believe that the results are not independent of the results of the PB-LC Study. Other recent studies conducted on different topics by different researchers could have been reported in the journals or newspapers that we read, but it's unlikely that the results of such studies are non-independent of the results of the PB-LC Study; hence they are not relevant. Studies from 100 years ago fail to be relevant even if there is reason to believe that their results are not independent of the results of the PB-LC Study, since they weren't candidates for us to learn about on this particular occasion.

The Triviality Problem simply evaporates; that problem derived from the strategy of *weakening* the evidence—i.e., from **The PB-LC Study yielded a significance value of .01** to the weaker **At least one of the thousand studies yielded a significance value of .01**. But, here, we're *strengthening* the evidence—from **The PB-LC Study yielded a significance value of .01** to the *stronger* **We're learning only that the PB-LC Study yielded a significance value of .01**. So there's no worry about weakening the evidence to the point of triviality.

Fourth, the Independence Problem also evaporates, since the whole point of the analysis above is that another study is relevant to the interpretation of the results of the PB-LC Study only if there's reason to believe that the results of that study are *not* independent of the results of the PB-LC Study.

Fifth, we don't have to take the researchers' private psychological states into account in an unacceptable way when we do judge the existence of other studies to be relevant to the evidential impact of the PB-LC Study. Only other *actual* studies are candidates for us to learn about; it doesn't matter if some researcher *would have* kept conducting studies if he didn't like the results of the PB-LC Study.

And sixth, there is no issue about non-commutativity of evidence when we do judge another study to be relevant. Once we fix which studies were conducted and which ones were candidates for us to learn about, the evidential impact of that set of studies will be the same regardless of the order in which they were conducted or in which we learn about them.

Still, I think that it's appropriate to regard the above only as a fairly modest vindication of Defeatism, for three reasons.

First, as already stressed, Defeatism is true only in (some) cases where we have reason to believe that the results of the relevant studies are non-independent and positively correlated. But even though it is implausible that we have this sort of reason *in general*, the original Defeatist intuitions and motivations don't seem to be sensitive to the existence of this sort of reason. Even supposing, for example, that we have some *guarantee* (from God, say) that any correlations between peanut butter and any of the health characteristics being considered are independent, one might still be tempted to think that the existence of the 999 Other Studies at least partially defeats the PB-LC Evidence. The appropriateness of applying the Bonferroni Correction, for example, doesn't depend on the agent having any independent reason to believe that the results of the studies are positively correlated. And the original Defeatist thought that *at least one of the studies was bound to yield a statistically significant correlation* still seems to apply, regardless of whether the results are positively correlated or not. If I'm right, then this thought is simply misguided.

Second, it should be clear that the Defeatist response to finding out about the existence of the 999 Other Studies is appropriate only for those of us who have "incomplete" access to the results of the 999 Other Studies—i.e., for those of us who are in either Situation 2 or Situation 3. But for the *researchers themselves*, who find out about the results of each study that they perform, Defeatism is false. Of course, on the assumption that they have reason to believe that the results of the thousand studies are positively correlated, they might have reason to alter their credence in the existence of a PB-LC Connection once they see the results of the 999 Other Studies; if all 999 Other Studies failed to yield any statistically significant results, for instance, the researchers would have reason to become less confident that the statistically significant results of the PB-LC Study are due to the existence of a PB-LC Connection (since the 999 Other Studies would be independent evidence against the existence of a PB-LC Connection). By the same taken, lots of statistically significant results among the 999 Other Studies would be independent evidence *for* the existence of a PB-LC Connection. But suppose that the number of statistically significant results among the 999 Other Studies is such that, taken as a whole, those results are neither independent evidence for nor independent evidence against the existence of a PB-LC Connection. In such a case, since neither the considerations from Situation 2 nor from Situation 3 apply, the Defeatist reaction is inappropriate for the researchers themselves, even if they do have independent reason to think that the results of the thousand studies are positively correlated. And this fact seem to me to limit the scope of the Defeatism argued for in this section considerably.

Third, I think that there's good reason to believe that the Defeatist effect in the peanut butter case is actually rather weak. If a researcher gets a statistically significant result from the PB-LC Study, it's unlikely that very much of the "competition" for our finding out about that result really comes from the other studies run by *that researcher* investigating possible peanut butter connections. Even if he runs the 999 Other Studies, most of the relevant "competition" really comes from *other* researchers conducting wholly unrelated studies that might be published in the journals and newspapers that we read. So, in real life cases like the peanut

butter case, I just don't think that the fact that PB-LC Study yielded a statistically significant result does very much at all to increase its own competition for being learned about by making it likelier that other peanut butter studies will also yield statistically significant results; thus, Situation 3–like effects will be rather weak. Similarly, I think it's fairly unlikely that Situation 2–like effects justify much of a Defeatist reaction to the peanut butter case. Do we really have an independent reason to think that the existence of a PB-LC Connection makes it likelier that there is also a PB-Gingivitis Connection? Reading through the side-effects of just about any drug on the market will convince one that a drug that is correlated with some positive effect is just as likely to be correlated with some negative effect as it is to be correlated with some other positive effect, and I can't imagine any reason to think that things are different with peanut butter.

7 Error Theory

In light of both the somewhat modest circumstances outlined in Section 6 in which Defeatism is true and the relative robustness of Defeatist intuitions, I think that some work remains to be done for an error theory; it would be nice to be able to explain away the Defeatist intuitions that many people (myself included) have in a way that is consistent with the falsity of Defeatism. In this section, I'll summarize three important such sources of Defeatist intuitions.

The first is that there are some pieces of information that we could learn which *seem* relevantly like the information that the 999 Other Studies were performed, and which would at least partially defeat the PB-LC Evidence. The first is the information that several other studies investigating a possible correlation *between peanut butter consumption and low cholesterol* were performed *and that none of them yielded a statistically significant correlation*. Everyone should agree that *this* information should make us less confident that there is a real PB-LC Connection; the failure of any particular study (and certainly of 999 studies!) to find a statistically significant correlation between peanut butter consumption and low cholesterol is obviously some evidence that no PB-LC Connection exists, regardless of the results of the PB-LC Study itself.

Moreover, even if we find out that 999 other studies investigating a PB-LC Connection were conducted and *don't* find out about the results of those studies, it's plausible that we are in a situation where Defeatism is true. After all, several studies investigating *the very same question* are clearly non-independent and positively correlated, and unless we had reason to believe that this *particular* PB-LC study (i.e., PB-LC Study #659, say) was chosen in advance to be the only PB-LC study that we would learn the results of (i.e., unless we had some reason to think that we were in Situation 1), it's plausible that we are in either Situation 2 or 3, in which case the Defeatist reaction is appropriate.

As discussed in Section 5, this case is relevantly like the fish net case above, where the fact that we were bound to catch a big fish completely defeated the evidential effect that the big fish we actually caught had on the hypothesis that the pond contains mostly big fish. The same point would apply to a version of

the pet monkey case from Section 2 where, instead of learning that I have one trillion pet monkeys that spend all day typing, you learn instead that I have made *George* type every day for one trillion days; in that case, I was bound to be able to produce putative evidence that *George* is far more intelligent than average, and the putative evidence that I do produce for that conclusion is defeated. But the differences between these cases and the original peanut butter case are quite subtle, and I think it's plausible that our intuitive reactions to the cases don't track these subtle differences precisely. As a result, I think that we often make the mistake of over-generalizing from cases where numerous studies are conducted investigating the very same question, and erroneously conclude that the same sort of evidential defeat is present in cases like the original peanut butter case.

The second source of Defeatist intuitions is that even if we accept the result that the PB-LC Evidence isn't defeated by the information that the 999 Other Studies were conducted, it doesn't follow that there aren't any important epistemic differences between the Single Study Scenario and the Multiple Study Scenario.

The reason is that, even if the information that the 999 Other Studies were conducted doesn't defeat the evidence that the PB-LC Study provides for the proposition that there's a PB-LC Connection, it still might defeat the evidence that the PB-LC Study provides for the proposition that there is a real connection between peanut butter and at least one of: low cholesterol, low self-esteem, low incidence of gingivitis, etc. This is somewhat counterintuitive. It's at least somewhat natural to think that, when you know that $H1$ entails $H2$, evidence for $H1$ is also evidence for $H2$; thus, evidence that the butler committed the murder is evidence that someone on the mansion staff committed the murder. Similarly, it's at least somewhat natural to think that, when you know that $H1$ entails $H2$, evidence against $H2$ is also evidence against $H1$; thus, evidence that nobody on the mansion staff committed the murder is evidence that the butler didn't commit the murder. But these principles are both false.[46,47]

Thus, even if the results of various studies are probabilistically independent, learning that you're in the Multiple Study Scenario and Situation 3 might well partially defeat the evidence that the PB-LC Study provides for the proposition that there is a real connection between peanut butter and at least one of: low cholesterol, low self-esteem, low incidence of gingivitis, etc. It might seem to follow from this that the evidence that the PB-LC Study provides for a PB-LC Connection has thereby also been defeated, which would give rise to the Defeatist intuition. But that intuition is mistaken.

Third, and finally, there are some conclusions about the *researchers themselves* that seem to be justified merely by learning something about their investigative practices.[48] If asked to give my credence right now that peanut butter consumption and low cholesterol are really connected, I suppose I would give a rather low number like .01 or .02. But if I were to learn that reputable researchers are conducting a study designed to look into that question, it might be reasonable for me to take that *all by itself* to be some evidence that there is a real PB-LC Connection, quite independently of what the researchers find. After all, I might reason, most reputable researchers don't just go around conducting studies that they don't have any reason

to believe might lead somewhere, so their conducting the PB-LC Study is some evidence that they do think it might lead somewhere, which is some (defeasible, of course) evidence that it will in fact lead somewhere. Moreover, *this* reason to believe that there is a PB-LC Connection seems to be in place only when the PB-LC Study is one of very few studies being performed; a plausible explanation for the researchers conducting the PB-LC Study along with perhaps one or two others is that they have some specific reason to think that there is a PB-LC Connection, whereas a much better explanation for their conducting *one thousand* studies is that they were just taking one thousand shots in the dark. So learning that the 999 Other Studies were conducted in addition to the PB-LC Study might be some evidence that the researchers *didn't* have any independent reason to think that there is a PB-LC Connection, and hence a reason to believe that there is no such correlation.

But it should be clear that this kind of defeat is fundamentally different from the sort of defeat that we have been considering so far. Epistemologists standardly distinguish between "undercutting" defeaters, which somehow undermine the probative force of some evidence with respect to a given hypothesis, and "opposing" defeaters, which provide positive reason to disbelieve the hypothesis.[49] The real interesting question about Defeatism seems to be the question of whether the existence of the 999 Other Tests serves as an *undercutting* defeater of the PB-LC Evidence; this effect looks to be an example of *opposing* defeat, where the fact that the 999 Other Studies were conducted *itself* gives us a reason to become less confident that there is a real PB-LC Connection, quite independent of the results of the PB-LC Study. Also, I think it's implausible that any real Defeatist effect here is particularly strong. Under ordinary circumstances where only one study is performed, I don't think that a hypothesis about the researcher's private reasons for conducting that study plays a crucial role in grounding the conclusions that we draw from the results of the study; after all, these reasons typically aren't cited in medical or scientific journals, and classical statisticians insist that we can evaluate the probative force of a study *without* having to know anything about independent evidence for or against some connection.

Relatedly, I don't think that we would be at all hesitant to become quite confident in the existence of a PB-LC Connection in a case where the PB-LC Study was the only one conducted but where the researcher herself claimed that she ran the study just because she was curious, or just to try to disprove an old wives' tale, or even because a dart that she threw at a dartboard full of health characteristics landed on the region marked "low cholesterol." But all of these explanations for why she conducted the PB-LC Study neutralize any reason to think that she had an independent justification to believe in a PB-LC Connection, and thus should make us much less confident in the existence of a PB-LC Connection if the Defeatist effect discussed above were playing a crucial role in our reasoning. Thus, I think it's implausible that learning that the researcher *lacked* an independent reason to suspect a PB-LC Connection (as we might learn when we find out that the 999 Other Studies were conducted) could have a significant impact on our final credence that there is a real PB-LC Connection.

8 Conclusion

What has been shown? I have been arguing that there are a variety of insurmountable obstacles to the naive application of the Defeatist position to all situations in which there are multiple tests or trials of some sort. The Defeatist point that since there were so many trials, *some* surprising outcome was bound to occur cannot have the general argumentative force that the Defeatist takes it to have. Defeatism is, however, the correct view under reasonably rare but precisely articulable circumstances, and its application in those circumstances avoids the obstacles to applying Defeatism generally. Still, this kernel of truth to Defeatism does only some of the work needed to explain the pervasiveness of Defeatist intuitions; I have tried to identify three further sources of Defeatist intuitions, each of which is consistent with anti-Defeatism. To the extent that arguments in science, statistics, and philosophy rely on a general application of the Defeatist view, they need to be re-evaluated.[50]

Notes

[1] This is crucial. Of course, we *might* reasonably take the performance of the thousand studies to be some evidence that the researchers have somehow deliberately falsified the study, in which case we obviously have less reason to believe in a real connection between peanut butter and low cholesterol. So I want to just assume that no *uncontroversial* cases of scientifically unethical behavior have taken place, and address the controversial (and far more interesting) question of whether the performance of the thousand studies *itself* does anything to defeat the results of the study under consideration.

[2] There may be other, non-causal explanations for non-chance correlations. For example, the fact that Samuel Clemens's height, weight, etc. are highly positively correlated with Mark Twain's height, weight, etc. is perhaps best explained by the *metaphysical* fact that Samuel Clemens and Mark Twain are identical, rather than by any *causal* fact. I mention this sort of case just to set it aside; the non-chance explanations we will be concerned with here will be causal ones.

[3] See Sober (ms) for an account of non-coincidence explanation in terms of the existence of some causal connection or other.

[4] Thanks to an anonymous referee from Noûs for pressing me to make this clarification.

[5] See, e.g., Dowe forthcoming, Hacking 1987, Juhl 2005, Manson and Thrush 2003, Parfit 1998, and White 2000.

[6] See, e.g., Achinstein 1994, Collins 1994, Harker 2006, Horwich 1982, Maher 1988, Schlesinger 1987, van Fraassen 1980, and White 2003.

[7] This is slightly complicated by the need to choose a so-called "test statistic" which determines the so-called "outcome space." Data that is significant at the α level, then, corresponds to an outcome such that, according to the null hypothesis, the probability that that outcome or some element of the outcome space at least as improbable would occur was α. However, this complication will not concern us here; see Howson and Urbach 1993 for discussion. Thanks to Jason Grossman for clarification here.

[8] To avoid irrelevant complications, I will also assume that the β of the test, which corresponds to the probability of *not* collecting the relevant data if there *is* a real connection between peanut butter and low cholesterol, is low. However, my focus will be on α, not β.

[9] McCabe and Moore 2003, p. 465.

[10] McCabe and Moore are actually considering a slightly different case than the one where multiple studies are conducted; they're considering a case where all of the data is collected at once, and multiple *statistical tests* are conducted on that data. But I think it's fairly clear that this is a distinction without a difference, and turns on the merely terminological issue of how to use the word "study." Whatever conclusions are justified in the Multiple Study Scenario in the peanut butter case, surely the same conclusions would be justified if the researchers simultaneously collected lots of data on people's peanut butter consumption, cholesterol levels, incidence of heart disease, incidence of gingivitis, etc., and then

ran one thousand significance tests on that body of data. Either way, the Defeatist thought that the researchers were bound to find some statistically significant correlation or other clearly applies.

[11] McCabe and Moore 2003, p. 465.

[12] See, e.g., Weiss 2004, pp. 822–827, Bluman 2006, p. 536, LeMoine 2004, p. 29, and Ender 1998.

[13] The Bonferroni Correction is actually an approximation of the Šidák Correction, which entails that the real "per family" significance level of a study is $1 - (1 - \alpha_{PT})^C$, where α_{PT} is the "per test" significance level that each individual test would have if it had been performed alone, and C is the number of tests that were performed. The Bonferroni Correction equation is the first linear term of the Taylor expansion of the Šidák Correction equation. In the vast majority of cases, the Bonferroni approximation is very close to the true Šidák value; the Bonferroni equation is used because it is easier to compute. The distinction between these two Corrections will not concern us here. See Abdi 2007 for a discussion.

[14] Abdi 2007, p. 103.

[15] See, e.g., Whitehead 1993, Gillies 1990, Howson and Urbach 1993, pp. 365–6, and Hacking 1965, pp. 107–9.

[16] However, I do think that much of what I say here about multiple studies carries over to the stopping rules case. See Kotzen ms b for further discussion.

[17] For proofs of the so-called "Infinite Monkey Theorem," see, e.g., Isaac 1995, pp. 48–50 and Gut 2005, pp. 97–100.

[18] See Feldman 1985, Feldman 1993, Conee and Feldman 1998, and Beebe 2004 for discussions of the Generality Problem for reliabilist views of justification. See Reichenbach 1949 and Hájek forthcoming for a discussion of the closely related Reference Class Problem that arises for various views on the nature of probabilities. See Levi 1977 and Kyburg 1977 for a discussion of the Reference Class Problem as it arises in cases of "direct inference."

[19] This is important, because it's straightforwardly possible for a test that is significant at some level when considered to be a token of one type to fail to be significant at that level when considered to be a token of a different type. In fact, the Defeatist's strategy *depends* on this being possible; otherwise, the fact that **At least one of the thousand studies yielded a significance level of .01** is very likely to be true in the Multiple Study Scenario would be irrelevant.

[20] One might think that the Defeatist is actually *strengthening* the evidence, from **The PB-LC Study yielded a significance value of .01** to **The PB-LC Study yielded a significance value of .01 and the 999 Other Studies were conducted**. But this would make the Defeatist strategy rather mysterious; though **At least one of the thousand studies yielded a significance level of .01** is very likely to be true in the Multiple Study Scenario, **The PB-LC Study yielded a significance value of .01 and the 999 Other Studies were conducted** is very *unlikely* to be true even in the Multiple Study Scenario; in fact, it's no more likely to be true than **The PB-LC Study yielded a significance value of .01**. So I think that the Defeatist really does have the strategy of weakening the evidence in mind.

[21] The requirement that we always update on the *strongest* piece of information that we know is sometimes referred to as **The Requirement of Total Evidence**. There are clear cases where violation of this Requirement leads to trouble. Suppose, for example, that there are two urns, each of which contains four marbles. Urn 1 contains two red marbles and two yellow marbles, and Urn 2 contains one red marble, one yellow marble, and two blue marbles. One of the urns is selected at random and a random marble is selected from it, which turns out to be red. This is evidence that the marble was drawn from Urn 1, since two out of the four marbles in Urn 1 are red, whereas only one out of the four marbles in Urn 2 is red. But a weaker way of describing the evidence is as **The randomly selected marble is either red or blue**. And if we update on this weaker information, we end up with a better reason to believe that the marble was drawn from Urn 2; after all, three out of the four marbles in Urn 2 are red or blue, whereas only two out of the four marbles in Urn 1 are red or blue. But this is clearly the wrong conclusion to draw. As a result, the Defeatist owes us an explanation of why *his* proposed weakening of the evidence is epistemically appropriate. See Sober ms for a discussion of this point.

[22] Perhaps there is some reason to think that the results of the thousand studies aren't completely independent. But I think that the Defeatist reasoning is just as intuitively plausible in cases where we have some *guarantee* (say, from God) that the results of the studies are independent. We'll investigate both the plausibility and the effect of relaxing this Independence assumption in Section 6.

[23] This example is from White 2000.

[24] This is also a problem for the classical methodology for calculating "P-values"; see Howson and Urbach 1993 and Kotzen ms b for a discussion.

[25] For instance, see Doring 1999, Field 1978, Kelly forthcoming, Skyrms 1986, van Fraassen 1989, and Weisberg ms. Many psychologists also accept the principle (see, e.g., Baron 2000, p. 197), as do many mathematicians (see, e.g., Diaconis and Zabell 1982 and Wagner 2002).

[26] Here's one complication: if I learn E before F, it seems as though that would make it rational for me to believe **I learned E before F**, whereas learning F before E would not make it rational to believe that claim. Here's another: observing the chair upright, then at a 45° angle to the ground, then on the ground might make it rational to believe **The chair just fell down**, whereas observing the reversed sequence might make it rational to believe instead **The chair just righted itself**. See Weisberg ms for discussion.

[27] See Kelly forthcoming for discussion.

[28] Thanks to Sinan Dogramaci for helpful discussions about this approach.

[29] Even though I'm stipulating that dice are being removed from the jar without replacement, I don't think that this affects any of the conclusions that I draw in this section. First of all, even if non-replacement affects the independence of Harry's outcome from the *outcomes* of the other rolls, it doesn't affect the independence of Harry's outcome from the *fact that 999 other dice were rolled*. Moreover, I've made the number of dice large in comparison to the number of dice drawn from the jar; if you'd like, we can make the number even larger so as to make the effect of non-replacement arbitrarily small. Or, we could stipulate that whenever a biased die is removed from the jar, God immediately replaces it with another biased die, and the same for fair dice.

[30] More precisely, the probability is $1 - (1 - (\frac{1}{6})^3)^{1000} \approx .99035$.

[31] Of course, he could deny this, and claim that it should be higher than .01 (and that, presumably, it should be .01 when 1,000 dice are drawn at the same time), but this seems unmotivated and obviously false.

[32] In a common form, that rule is that where p is an agent's credence function before learning E, his new credence p_{new} in any hypothesis H after learning E should be $p_{new}(\mathrm{H}) = p(\mathrm{H} \mid \mathrm{E}) = \frac{p(H) \times p(E \mid H)}{p(H) \times p(E \mid H) + (1 - p(H)) \times p(E \mid \neg H)}$. Thus, all we need to calculate $p_{new}(\mathrm{H})$ according to this rule is $p(\mathrm{H})$, $p(\mathrm{E} \mid \mathrm{H})$, and $p(\mathrm{E} \mid \neg\, \mathrm{H})$.

[33] See, e.g., Christensen 2004, Earman 1992, Greaves and Wallace 2006, Howson and Urbach 1993, Joyce 1998, Kaplan 1996, Maher 1993, and Sober 2002.

[34] For instance, I'm not making essential use of any of the more controversial parts of the Bayesian program—for instance, the assumption of logical omniscience or the irrelevance of "old evidence." All I'm assuming is that when we're in an "ordinary" situation with well-defined priors and likelihoods, we should update our credences by Conditionalization.

[35] There are actually two problems for the Defeatist here. First, since he updates in a manner other than by Conditionalization, he can be "Dutch Booked"—i.e., he'll be disposed to accept each of a series of bets which is guaranteed to lose money as a package. See Earman 1992 and Howson and Urbach 1993 for Dutch Book proofs and discussion. Second, a straightforward "Monte Carlo" simulation demonstrates that the Defeatist should expect to lose money in the long run by repeatedly betting in accordance with his credences, even without the need for a clever bookie. See Kotzen ms b for discussion.

[36] $p(\textbf{HarryIsBiased} \mid \textbf{Three6s}) = \frac{p(\textbf{HarryIsBiased}) \times p(\textbf{Three6s} \mid \textbf{HarryIsBiased})}{p(\textbf{HarryIsBiased}) \times p(\textbf{Three6s} \mid \textbf{HarryIsBiased}) + p(\neg\textbf{HarryIsBiased}) \times p(\textbf{Three6s} \mid \neg\textbf{HarryIsBiased})} = \frac{(.01)(1)}{(.01)(1) + (.99)(\frac{1}{6})^3} \approx .68571$.

[37] One might worry that a crucial disanalogy between the peanut butter case and the dice case is that our knowledge of the composition of the jar gives reason to have a particular credence in **HarryIsBiased**, whereas we have no idea what percentage of health characteristics peanut butter is really connected to. Still, surely we have *some* nonzero credence in the existence of a PB-C Connection before reading the PB-C Study, and I can't see what difference it could make what that credence is based on.

[38] This example derives from Eddington 1939.

[39] Let's assume, for simplicity, that this is true—i.e., that the circumstances were such that we were bound to catch a large fish in our net regardless of the ratio of large fish to small fish in the pond.

[40] The most famous example of an evidential distinction between **p** and **the agent is learning that p** is the so-called "Monte Hall Problem." See, e.g., Bapeswara and Rao 1992 and Gill 2002 for discussion.

[41] Thanks to Roger White for pressing me on the relevance of Situation 2.

[42] In other words, that a real PB-Gingivitis Connection (for example) is likelier on the assumption of a real PB-LC Connection (and vice versa)

[43] Obviously, there is a spectrum of possibilities here; the researchers could choose, for example, to randomly select two, or three, etc., of the statistically significant studies to tell us about. But it will be useful to focus on the extreme case where we only learn about one of the studies; all of my claims will apply equally to cases where we learn about any proper subset of the statistically significant results.

[44] To make this more concrete, suppose that we have two coins which have positively correlated biases; our credence is .5 that they're both fair, and our credence is .5 that they're both biased so as to land heads $\frac{2}{3}$ of the time. If we learn that coin A landed heads, our posterior credence that A is biased is: $p(\text{ABiased} \mid \text{AHeads}) = \frac{(\frac{1}{2})(\frac{2}{3})}{(\frac{1}{2})(\frac{2}{3})+(\frac{1}{2})(\frac{1}{2})} = \frac{4}{7}$. But suppose that we know that both coins will be flipped, that we will learn of how a coin landed only if it landed heads, and that if both coins land heads, then we will learn of only one outcome, selected at random. Then (where 'N' is an abbreviation for the numerator of the fraction), $p(\text{ABiased} \mid \text{ToldAHeads}) = \frac{(\frac{1}{2})[\frac{2}{3}((\frac{1}{3})+(\frac{2}{3})(\frac{1}{2}))]}{N+(\frac{1}{2})[(\frac{1}{2})((\frac{1}{2})+(\frac{1}{2})(\frac{1}{2}))]} = \frac{32}{59} < \frac{4}{7}$, so there is defeat of the evidence that we have acquired for ABiased.

[45] And, of course, the reasoning above applies to *any* study whose results we have reason to believe are non-independent of the results of the PB-LC Study.

[46] See Kotzen ms a for a discussion of the former principle.

[47] To see a counterexample to the latter principle, suppose that you have two coins, A and B, and that you have an independent credence of .5 that each coin is fair, and an independent credence of .5 that each coin is $\frac{2}{3}$ heads-biased.

Suppose first that you're in the analog of Situation 1 from Section 6: Coin A will be flipped, and you're going to find out about how A lands regardless of what happens with Coin B. When you learn that coin A landed heads (AHeads), your updated credence in the proposition that coin A is $\frac{2}{3}$ heads-biased (ABiased) is:

$$p(\text{ABiased} \mid \text{AHeads}) = \frac{(\frac{1}{2})(\frac{2}{3})}{(\frac{1}{2})(\frac{2}{3})+(\frac{1}{2})(\frac{1}{2})} = \frac{4}{7}.$$

Since AHeads doesn't tell you anything about Coin B, your credence that Coin B is biased is still $\frac{1}{2}$, so your credence that at least one of the coins is biased (SomeBiased), is:

$$p(\text{SomeBiased} \mid \text{AHeads}) = \frac{1}{2} + \frac{4}{7} - \frac{4}{14} = \frac{11}{14} \approx .7857.$$

Now, suppose that you were to find out that you're in the analog of Situation 3 from Section 6: Both coins were flipped, you're told about the outcome of a flip only if it lands heads, and you're told about at most one outcome (if both coins land heads, one will be selected at random for you to learn about). This does nothing to defeat your credence in ABiased (to save space, I've used 'N' to denote the numerator of the fraction):

$$p(\text{ABiased} \mid \text{ToldAHeads}) = \frac{(\frac{1}{2})\{(\frac{2}{3})[(\frac{7}{12})(\frac{1}{2})+(\frac{5}{12})(1)]\}}{N+(\frac{1}{2})\{(\frac{1}{2})[(\frac{7}{12})(\frac{1}{2})+(\frac{5}{12})(1)]\}} = \frac{4}{7}.$$

But this does partially defeat your credence in SomeBiased:

$$p(\text{SomeBiased} \mid \text{ToldAHeads}) =$$
$$\frac{(\frac{3}{4})\{(\frac{1}{3})(\frac{2}{3})[(\frac{1}{2})(1)+(\frac{1}{2})(\frac{1}{2})]+(\frac{1}{3})(\frac{1}{2})[(\frac{1}{3})(1)+(\frac{2}{3})(\frac{1}{2})]+(\frac{1}{3})(\frac{2}{3})[(\frac{1}{3})(1)+(\frac{2}{3})(\frac{1}{2})]\}}{N+(\frac{1}{4})[(\frac{1}{2})(\frac{1}{2})+(\frac{1}{2})(\frac{1}{2})(\frac{1}{2})]} \approx .7731.$$

So, even if the results of the coin flips are probabilistically independent, the information that you're in the analog of Situation 3 can partially defeat your credence in SomeBiased without partially defeating your credence in ABiased, even though you know that ABiased entails SomeBiased.

[48] See White 2003 for a discussion of inferences about the *researchers themselves* as opposed to the *theories* that they produce.

[49] For example, if Mary's testimony that it is raining out is evidence that it is raining out, John's testimony that it's not raining out is an opposing defeater, whereas Kate's testimony that Mary is very unreliable at reporting the weather is an undercutting defeater. See Kotzen ms c for a discussion and formal account of this distinction.

[50] Thanks to Yoav Benjamini, Eliza Block, Paul Boghossian, Shamik Dasgupta, Sinan Dogramaci, Adam Elga, Dana Evan, Hartry Field, Kit Fine, Alexis Gallagher, Don Garrett, Jason Grossman, Terry McGovern, John Morrison, Tom Nagel, Jill North, Jim Pryor, Karl Schafer, Josh Schechter, Stephen Schiffer, Juliet Shaffer, Michael Strevens, Peter Unger, and Roger White for helpful comments on earlier drafts of this paper.

References

Abdi, H. (2007). "Bonferroni and Šidák corrections for multiple comparisons," in N.J. Salkind (ed.): *Encyclopedia of Measurement and Statistics*, pp. 103–7. Thousand Oaks, CA: Sage.

Achinstein, P. (1994). "Explanation v. Prediction: Which Carries More Weight?," in Hull, Forbes, and Burian (eds) 1994, *Proceedings of the Philosophy of Science Association* Vol. 2. East Lansing, Mich.: Philosophy of Science Association, pp. 156–64.

Bapeswara, V. and Rao, B. (1992). "A three-door game show and some of its variants," *The Mathematical Scientist* 17(2), pp. 89–94.

Baron, J. (2000). *Thinking and Deciding*. Cambridge: Cambridge University Press.

Beebe, J. (2004). "The Generality Problem, Statistical Relevance and the Tri-Level Hypothesis," *Noûs* 38(1), pp. 177–195.

Bluman, A. (2006). *Elementary Statistics, 3rd Edition*. McGraw-Hill.

Christensen, D. (2004). *Putting Logic in its Place: Formal Constraints on Rational Belief*. Oxford: Clarendon Press.

Collins, R. (1994). "Against the Epistemic Value of Prediction over Accommodation," *Noûs* 28, pp. 210–24.

Conee, E. and Feldman, R. (1998). "The Generality Problem for Reliabilism," *Philosophical Studies* 89, pp. 1–29.

Diaconis, P. and Zabell, S. (1982). "Updating Subjective Probability," *Journal of the American Statistical Association* 77(380), pp. 822–830.

Doring, F. (1999). "Why Bayesian Psychology Is Incomplete," *Philosophy of Science*, Vol. 66, Supplement. *Proceedings of the 1998 Biennial Meetings of the Philosophy of Science Association. Part I: Contributed Papers* (Sep., 1999), pp. S379–S389.

Dowe, P. (forthcoming). "The Inverse Gambler's Fallacy Revisited: Multiple Universe Explanation of Fine Tuning."

Earman, J. (1992). *Bayes or Bust?*. Cambridge: MIT Press.

Ender, P. (1998). Notes from Education 230A: Introduction to Research Design and Statistics, available at http://www.gseis.ucla.edu/courses/ed230a2/notes2/logic.html.

Eddington, A. (1939). *The Philosophy of Physical Science*. Cambridge: Cambridge University Press.

Feldman, R. (1985). "Reliability and Justification," *The Monist* 68, pp. 159–174.

Feldman, R. (1993). "Proper Functionalism," *Noûs* 27, pp. 34–50.

Field, H. (1978). "A Note on Jeffrey Conditionalization," *Philosophy of Science* 45(3), pp. 361–367.

Gill, J. (2002). *Bayesian Methods*. CRC Press.

Gillies, D. (1990). "Bayesianism versus Falsificationism," *Ratio (New Series)* III(1), pp. 82–98.

Greaves, H. and Wallace, D. "Justifying conditionalization: Conditionalization maximizes expected epistemic utility," *Mind* 115, pp. 607–632.

Gut, A. (2005). *Probability: A Graduate Course*. Springer.

Hacking, I. (1965). *Logic of Statistical Inference*. London: Cambridge University Press.

Hacking, I. (1987). "The Inverse Gambler's Fallacy: the Argument from Design. The Anthropic Principle Applied to Wheeler Universes," *Mind* 76, pp. 331–340.

Hájek, A. (forthcoming). "The Reference Class Problem is Your Problem Too," forthcoming in *Synthese*.
Harker, D. (2006). "Accommodation and Prediction: The Case of the Persistent Head," *British Journal for the Philosophy of Science* 57(2), pp. 309–321.
Horwich, P. (1982). *Probability and Evidence*. Cambridge: Cambridge University Press.
Howson, C. and Urbach, P. (1993) *Scientific Reasoning: The Bayesian Approach, 2nd Edition*. Chicago: Open Court.
Isaac, R. (1995). *The Pleasures of Probability*. Springer.
Joyce, J. (1998). "A Nonpragmatic Vindication of Probabilism," *Philosophy of Science* 65, pp. 575–603.
Juhl, C. (2005). "Fine tuning, Many Worlds, and the 'Inverse Gambler's Fallacy,'" *Noûs* 39(2), pp. 337–347.
Kaplan, M. (1996). *Decision Theory as Philosophy*. Cambridge: Cambridge University Press.
Kelly, T. (forthcoming). "Disagreement, Dogmatism, and Belief Polarization," forthcoming in *Journal of Philosophy*.
Kotzen, M. (manuscript). "Dragging and Confirming." Chapter of Ph.D. Thesis.
Kotzen, M. (manuscript). "Stopping Rules and Evidential Defeat." Chapter of Ph.D. Thesis.
Kotzen, M. (manuscript). "A Formal Account of Defeat." Chapter of Ph.D. Thesis.
Kyburg, H. (1977). "Randomness and the Right Reference Class," *Journal of Philosophy* 74, pp. 501–21.
LeMoine, V. (2004). Notes from STAT 201: Elementary Statistical Inference, available at http://www.stat.tamu.edu/~vlemoine/201/lecture/chapter6.pdf.
Levi, I. (1977). "Direct Inference," *Journal of Philosophy* 74, pp. 5–29.
Maher, P. (1988). "Prediction, Accommodation, and the Logic of Discovery" in Fine and Leplin (eds) 1988, *Proceedings of the Philosophy of Science Association Vol. 1*. East Lansing Mich.: Philosophy of Science Association.
Maher, P. (1993). *Betting on Theories*. Cambridge: Cambridge University Press.
Manson, N. and Thrush, M. (2003). "Fine-Tuning, Multiple Universes, and the 'This Universe' Objection," *Pacific Philosophical Quarterly* 84(1), pp. 67–83.
McCabe, G. and Moore, D. (2003). *Introduction to the Practice of Statistics, 4th Edition*. New York: W.H. Freeman and Company.
Parfit, D. (1998). "Why anything? Why this?," *London Review of Books* Jan 22, pp. 24–27.
Pryor, J. (2004). "What's Wrong with Moore's Argument?" *Philosophical Issues* 14(1), pp. 349–378.
Reichenbach, H. (1949). *The Theory of Probability*. Berkeley: University of California Press.
Schlesinger, G. (1987). "Accommodation and Prediction," *Australasian Journal of Philosophy* 65, pp. 28–42.
Skyrms, B. (1986). *Choice and Chance, 3rd Edition*. Belmont, CA: Wadsworth.
Sober, E. (manuscript). "Coincidences and How to Reason about Them," available at http://philosophy.wisc.edu/sober/Coincidences%20no%20sids.pdf.
Sober, E. (2002). "Bayesianism—Its Scope and Limits," in Richard Swinburne, ed., *Bayes's Theorem* (Oxford: Oxford University Press): pp. 21–38.
van Fraassen, B. (1980). *The Scientific Image*. Oxford: Oxford University Press.
van Fraassen, B. (1999). *Laws and Symmetry*. Oxford: Clarendon Press.
Wagner, C. (2002). "Probability Kinematics And Commutativity," *Philosophy of Science* 69, pp. 266–278.
Weisberg, Jonathan (manuscript). "Commutativity or Holism?," available at http://www.utm.utoronto.ca/~weisber3/docs/JCv2.pdf.
Weiss, N. (2004). *Introductory Statistics, 7th Edition*. Addison Wesley.
White, R. (2000). "Fine-Tuning and Multiple Universes," *Noûs* 34, pp. 260–76.
White, R. (2003). "The Epistemic Advantage of Prediction Over Accomodation," *Mind* 112(448), pp. 653–683.
Whitehead, J. (1993). "The Case for Frequentism in Clinical Trials," *Statistics in Medicine* 12(15–16), pp. 1405–1413.