

## Am I Doing Bad Science?

Margaret Min

**Abstract:** The legitimacy of a scientific theory seems to depend heavily on whether or not that theory has made accurate predictions. For example, consider the theory that claims the misuse of antibiotics, like penicillin, can lead to potentially deadly antibiotic resistance. This theory was based on observations of the development of antibiotic resistance in bacteria, yet there was resistance to this theory until the theory proved itself many times over with confirmed predictions (i.e. many predicted instances of strains of bacteria evolving resistance to antibiotics). Today, the theory that misuse of antibiotics can lead to antibiotic resistance is widely accepted in the scientific community. More generally, most scientific theories are judged at least in part on whether they have predicted a large amount of data, to the point where it is necessary for theories to have predicted data to be taken seriously at all. Interestingly, pre-existing data that is explained by a new theory is typically viewed as less supportive of that theory than data that has been predicted. This paper will critique the widespread belief that predicted data is better than pre-existing explained data by reviewing two philosophers' opinions on this matter, Peter Lipton and Carl Hempel.

## I. What are we arguing about?

This paper will address two of the arguments Peter Lipton presents in his paper, “Testing Hypotheses: Prediction and Prejudice”. Before presenting those arguments, let me first explain the background information related to Lipton’s arguments. In postulating scientific hypotheses, there are two main types of data used. In this paper I will call them accommodated data and predicted data.

Accommodated data are data that are collected before a hypothesis is proposed. Typically, scientists will construct scientific hypotheses based on their interpretations of accommodated data. For example, a botanist who is battling an infestation of ants in his greenhouse using bug-spray might notice that the plants on which he sprays bug-spray tend to produce fewer seeds. The botanist observes 100 of his plants, 67 of which were sprayed with bug-spray and 33 of which were left alone. All 67 of the bug-sprayed plants produced fewer than half the number of seeds produced by the 33 non-bug-sprayed plants. From this (accommodated) data, the botanist proposes his scientific hypothesis that bug-spray has toxic effects on plants and makes them produce fewer seeds than normal. (Importantly, accommodated data is labeled as “accommodated” only for the role it plays in postulating scientific hypotheses. The hypothesis attempts to explain and account for the data, in other words, accommodate the data. However, nothing about

accommodated data taken alone should suggest that it is being accommodated or accommodating something else.)

Predicted data are data that are collected after a hypothesis is proposed. After the botanist proposes his hypothesis, he predicts that if he were to plant the seeds of the bug-sprayed plants in a no-bug-spray-zone, effectively eliminating the bug-spray factor, normal seed production would be restored in the next generation. He carries out this experiment and observes that his prediction was correct. The predicted data are his observations of the results of his experiment.

Generalizing to all scientific hypotheses, we see that all scientific hypotheses are supported by accommodated data (that is how hypotheses get proposed in the first place), while a majority of those scientific hypotheses are additionally supported by predicted data. A small subset of scientific hypotheses is supported only by accommodated data and not by predicted data (presumably, these would be very young hypotheses that have yet to be tested by experiments).

It is a widely accepted view that predicted data provide stronger support for hypotheses than do accommodated data. Lipton agrees with this intuition and argues for two versions of this view, what he calls the weak advantage thesis and the strong advantage thesis. In this paper, I will present Lipton’s weak advantage thesis (II) and explain my agreement with

it (III), followed by a presentation of Lipton's strong advantage thesis (IV) and my disagreement with it (V).

## II. Lipton's weak advantage thesis

One of the virtues of scientific hypotheses is having precise and relevant supporting data. This is an evidential virtue, meaning that it is a virtue of a hypothesis in nature of how well it is supported by empirical evidence. Peter Lipton, along with many scientists I take it, would agree that having precise and relevant supporting data is very important for scientific theories. A hypothesis that has weak supporting data is a worse scientific theory than a hypothesis that has strong supporting data, all other things being equal. Although Lipton does not go into too much detail in his paper about why he believes this is an evidential virtue, I would guess that he believes it in a simple and intuitive way as follows. Consider the botanist's hypothesis in Part 1 that states that bug-spray has toxic effects on plants and makes them produce fewer seeds. And let's say that instead of his actual data (observations of plants sprayed by bug-spray or not sprayed), the botanist instead collects his data from observations of goats that were sprayed with bug-spray. He observes that goats sprayed with bug-spray produce fewer offspring than goats that were left alone. Though this is an interesting phenomenon, observations of goats certainly provide less support for his hypothesis about plant seed production than would have observations

of actual plants. And all other things being equal, for any given hypothesis, we should prefer more strongly supportive data to weakly supportive data because this preference will yield better scientific theories.

Considering this fact, that a hypothesis is more strongly supported by precise and relevant data, Lipton presents his weak advantage thesis as follows:

Weak advantage thesis: Predictions tend to provide stronger support for hypotheses than do accommodations.

Lipton believes the weak advantage thesis because successful predictions tend to be more precise and relevant data than do accommodated data. Lipton argues that in typical scientific practice, a shrewd scientist will make observations on general phenomena and pick out just one or a few statistically significant patterns (i.e. a pattern that does not occur merely by chance) from the vast data. From these patterns, the scientists will make a hypothesis. With the hypothesis, the scientists will create experiments to further test the hypothesis. Importantly, these experiments are designed specifically to test the hypothesis in question, so the data collected (predicted data) will directly prove or disprove the hypothesis. In contrast, the original observed phenomena used to postulate the hypothesis (accommodated data) were not designed to answer any specific questions and much of the observed data may in fact be completely irrelevant to the resulting hypothesis except for select data points. Thus, Lipton

argues, predicted data tend to provide more powerful support for hypotheses than accommodated data because the way in which that data is collected. The passive way of collecting accommodated data can result in too much noisy and irrelevant data while the aggressive and specific experimentation done to collect predicted data will yield more useful and telling results.

Importantly, Lipton recognizes that these patterns of predictions providing stronger support than accommodations are not going to show up in every scientific theory. There may be cases in which the accommodated data is more supportive of the theory than predicted data. Perhaps the accommodated data came from observations of some phenomenon that cannot be repeated in artificial experiments. All the scientist can do is create models of the original phenomena. In this case, it seems less clear that the predicted data collected from these model experiments provides stronger support than the accommodated data collected from observations of the original phenomenon. These types of situations can occur, but Lipton claims only that predictions tend to provide stronger support than accommodations, and crucially not that predictions always provide stronger support than accommodations.

### **III. Agreement with the weak advantage thesis**

I agree with Lipton's weak advantage thesis. To illustrate, let's return to the botanist from Part I. Let's

tweak the situation a little and say that the botanist attempted to control his ant problem by spraying bug-spray on the plants, lowering the temperature of the greenhouse (ants prefer warm temperatures), and caressing the leaves of the plants to encourage them to stay strong in the face of the ant armies. The botanist records which plants he sprays bug-spray on, but does not record the plants he caresses or the temperature change in the greenhouse. Still, he makes his observations of 100 plants and sees that the 67 plants he sprayed bug-spray on produced fewer seeds than the non bug-spray plants. Using this data, he postulates the same hypothesis and designs further experiments to test it. Realizing that there are other potential variables (greenhouse temperature and relative botanist love/caressing), he designs further experiments to control for these variables to test only the effect of bug-spray. And what do you know his predictions are still correct! Bug-spray is the culprit of lower seed production, not lower temperatures or his excessive plant love. With this example, we can see that the accommodated data was certainly enough for the botanist to propose his hypothesis, but it was considerably weaker in support than the predicted data. The accommodated data was weaker simply because it was too unfocused, there were too many other factors that could have influenced the data. The predicted data was collected in a way that was designed specifically to eliminate those factors so

that the hypothesis at hand could be answered more definitively.

#### **IV. Lipton's strong advantage thesis**

Lipton's strong advantage thesis appeals to a different sort of virtue of scientific hypotheses, the theoretical virtue of simplicity. Ockham's razor is a general principle of theory selection that says that simple theories should be chosen over more complex theories (all other things being equal) because simpler theories are more likely to be true. This paper will not delve too deeply into the question of why simplicity plays a factor in choosing scientific theories (that would constitute its own paper!) and will instead present a simple example to show why Ockham's razor seems to be true in at least one way. Consider two explanations for why humans have 10 fingers. One explanation is that fingers evolved through 200 genetic mutations occurring over the last 100 million years. Another explanation is that fingers evolved through 2,000 genetic mutations occurring over the last 100 million years. Ockham's razor says that we should choose the first hypothesis because it is simpler - it requires 1,800 fewer genetic mutations to reach the same outcome. Considering the fact that genetic mutations are very rare, and genetic mutations specifically towards finger evolution are even rarer, the first hypothesis, statistically speaking, is more likely. The first hypothesis requires fewer really rare occurrences to occur. More generally, theories

that minimize the need for really rare occurrences to occur are simpler and more likely. This is only one of many types of simplicity. Some types of simplicity can conflict with each other, others are difficult to judge, and still more are hard to weigh against other types of simplicity.

Another virtue of scientific theories that plays a part in Lipton's second argument for the strong advantage thesis is the evidential virtue of having wide and diverse data supporting the hypothesis. To illustrate, consider a scientific hypothesis stating that the more chickens my grandmother raises, the more frequently she will eat omelets for breakfast. And let's say that the scientist that postulates this hypothesis has 2 data points to support it. When my grandmother has 15 chickens, she eats omelets 5/7 days of the week. When my grandmother has 16 chickens, she eats omelets 6/7 days of the week. The scientist reasons that the data points show that when we add one more chicken to my grandmother's family, she eats one more omelet every week and thus, the more chickens my grandmother has, the more omelets she will eat. However, the scientist's hypothesis is poorly supported because two data points is not enough to prove that more chickens means more omelets. For one, statistically speaking, we have no idea if these data points actually show a trend, or are just a result of chance fluctuations. Secondly, there may be other patterns of chickens and omelets that cannot be revealed by only 2 data points. For instance,

it could be that if my grandmother has fewer than 10 chickens, she will not eat any omelets, if she has 10-20 chickens, she'll eat 5 or 6 omelets a week, but any more than 20 chickens and she'll only eat fried chicken (i.e. no omelets). A trend like that can only be revealed by testing my grandmother extensively in each of those ranges and will not be discernible from a data set with only 2 points. Generally, having wide and diverse data will provide greater support for a hypothesis because it will more likely reveal consistent trends and thus lead to a hypothesis that is more accurate.

Lipton claims that there is a conflict between the theoretical virtue of simplicity and the evidential virtue of having wide and diverse data. In science, the ideal is to have both wide and diverse data, but also to have a simple theory. This can be challenging for scientists because the wider and more diverse their data is, the harder it is to keep the theory simple while still accounting for all that data. Thus, scientists are often required to make compromises between the two virtues, partly sacrificing a very simple theory as well as not being able to account for all the data at hand to meet at some happy middle. Lipton argues however, that scientists, being imperfect humans, are not very good at judging how much simplicity to sacrifice and how much data to sacrifice. Most scientists tend to try to keep all the data they can ("over-fitting" their data), sacrificing the simplicity of their hypotheses. Lipton calls this "fudging a hypothesis". Thus, many scientific

hypotheses are more complex than they need to be, and by Ockham's razor, they are thus less likely to be true than a simpler hypothesis. This brings us to Lipton's 2nd argument, what I will call his fudging argument, which supports the strong advantage thesis as follows.

Strong advantage thesis: A single, particular observation that was accommodated would have provided more support for the hypothesis in question if it had been predicted instead.

Why does Lipton think the strong advantage thesis is true? Lipton thinks that it is important to avoid making the common mistake of fudging a hypothesis (i.e. trying to explain too much of the data), for if we are able to avoid this mistake, we will have a simpler and thus more accurate theory. Predicted data arrive on the scene after a hypothesis or theory is already postulated. Thus, a scientist cannot fudge her hypothesis to fit more of her predicted data because the predicted data didn't exist when she was making the hypothesis. She wouldn't even know how to fudge her hypothesis to over-fit the predicted data. The scientist will only make the mistake when it comes to accommodated data, for that is the data used to create hypotheses. Thus, scientists will always run the risk of fudging their hypotheses since all scientific hypotheses are based in part on accommodated data. Scientists that have more accommodated data will be at a higher risk of fudging their hypotheses because there is more data available that they might want to explain by fudging hypotheses.

The strong advantage thesis is true because if the same hypothesis has less accommodated data than the other, but on total the same amount of supportive data (accommodated data + predicted data), then it is more likely to be a more accurate hypothesis because the scientist postulating that hypothesis had a lower risk of fudging it.

## V. Disagreement with the strong advantage thesis

### i) Disagreement #1

Lipton's support of the strong advantage thesis stems from his view that science should make a compromise between postulating a simple theory and accommodating for the most data they can, but human scientists are prone to imperfection at achieving this balance. Though I do agree that simplicity plays an important part in science, it seems that science should first prioritize explaining data. Science is in the business of explaining what the world is like and why the world is like that. And as humans, our only access to the world is through our experiences. In science, these experiences translate into our data and become our only resource in making any claims about the world. If scientists want to do good science, they should account for their data.

However, Lipton argues that we should actually sacrifice some of our data in the interests of simplicity. He says that data can often times be too wide and varied

to be accounted for in a simple theory. I would argue that though simplicity should still be kept in mind (for a very complex theory would be useless to us), the first priority in science is to explain the data. Simplicity should only come into play when comparing two hypotheses that explain the data equally well. Lipton calls it fudging hypotheses to explain more data, but his process of fudging hypotheses actually seems like good scientific practice. If a scientist has to trash some of her data because it doesn't agree with her simple theory, then she just hasn't found the best explanation for the data.

Crucially, I do not support the over-fitting of data. Instead, I value good statistical analyses that account for chance fluctuations in the data. When scientists appear to over-fit their data, it is not because they are trying to explain too much of the data in their theory. The problem instead lies in the fact that they have not used statistical analyses to explain data points that seem to deviate from trends. A scientist that employs effective statistical analyses will be able to explain all of her data while avoiding over-fitting.

Lipton could argue in response that even the use of statistical analyses is subject to the imperfection of humans, so a scientist that does statistical analyses will still run the risk of over-fitting her data if she does her statistical analyses inadequately. And because she runs this risk, it's ultimately better to just have more predicted data and less accommodated data since the

former will not pose this risk of over-fitting. I would argue in response that a scientist who does inadequate statistical analyses would find both accommodated data and predicted data to be problematic for a theory. On the one hand, doing inadequate statistical analyses on accommodated data may result in a hypothesis that over-fits the data, and that's obviously bad. But additionally, if the scientist is doing inadequate statistical analyses on predicted data, then she will likely make flawed interpretations of her data. She might take a certain data point that deviates from her hypothesis as evidence to reject her hypothesis, even if that data point was just a chance fluctuation in the experiment that could have been accounted for in statistically analyses. Or, she might look at her scattered data points and conclude that the empirical evidence supports her hypothesis, even if statistical analyses would show that the data actually indicates a different pattern. Thus, the fact that humans can be imperfect in carrying out statistical analyses poses a problem for scientific practice on the whole including both accommodated data and predicted data. The problem isn't when the data is collected, but that the scientist just doesn't know how to interpret the data. And so this human imperfection should not be a reason for us to prefer predicted data to accommodated data.

ii) Disagreement #2

It seems that even if Lipton's reasoning of fudging hypotheses were good reasoning, it would still fail

to support the strong advantage thesis. The strong advantage thesis focuses on how well a particular piece of data supports a hypothesis, that support depending on whether the data was found before or after the hypothesis is postulated. Lipton argues that the strong advantage thesis is true because hypotheses supported by fewer accommodated data are less likely to be fudged. His reasoning focuses on the chances an imperfect human scientist will fudge his hypothesis, and crucially does not focus on how supportive a piece of data is if it is accommodated or predicted. The contrast between accommodated and predicted data is revealed only coincidentally by how the two types of data happen to influence the likelihood of fudging. If we could build a machine whose only function is to balance simplicity and explaining the provided data to the exact levels Lipton would want (eliminating human imperfection), it seems like his fudging argument would no longer apply. However, the strong advantage thesis would still apply because hypotheses generated by this machine would be supported by varying amounts of accommodated data and predicted data. If Lipton wants to support the strong advantage thesis, then he will need to come up with an argument that directly addresses the difference in supportiveness of accommodated data and predicted data.

iii) Disagreement #3

Lipton's support of the strong advantage thesis relies on the imperfection of humans in being able

to construct simple theories that also accommodate for most of the data. I would like to focus on the imperfection of humans. My disagreement will first appeal to an argument made by Carl Hempel in his book *Philosophy of Natural Science*. Hempel says, "...the strength of the support that a hypothesis receives from a given body of data should depend only on what the hypothesis asserts and what the data are." In contrast, Lipton seems to argue in his fudging argument that the strength of support a given piece of data provides for its hypothesis depends on the likelihood that the imperfect human scientist has fudged his hypothesis to fit that piece of data. And Hempel (and I) would argue that that seems wrong. Science should be objective and certainly, the relation between theories and data should be objective. For any given hypothesis, how well it is supported by data should not depend on who proposed the hypothesis and how good they are at doing good science. Let's say Mary is really great at doing science (in at least the way Lipton thinks), but Jane has really bad judgment and often compromises simplicity to over-fit her data. Given a hypothesis A, supported by data points 1-10, how strongly those data points support hypothesis A should not depend on whether Jane or Mary is the scientist collecting the data and postulating the hypothesis. The "who" seems irrelevant to science. The fudging argument does not seem to be consistent with the objective nature of science.

Hempel (and I) would object to Lipton's fudging argument and his support of the strong advantage thesis in another way as follows. The differential valuing of predicted and accommodated data is misguided. Both types of data can provide equal support to any given hypothesis. How much support a piece of data can provide depends only on what the hypothesis asserts and what the data is, and crucially, does not depend on when the piece of data is collected. Hempel says that one way to illustrate this is to consider the spectral emission lines produced by certain gases, a phenomenon called the Balmer series. These emission patterns are described by a mathematically elegant and simple formula constructed by J. J. Balmer. Hempel would argue that when it comes to mathematically elegant hypotheses like this one, Lipton's fudging argument does not apply for there is no conflict involving simplicity, as the hypothesis is already very simple. Mathematically elegant hypotheses like this one are not over-fitting their data and are not fudged, even if those hypotheses happen to be constructed from lots of accommodated data. Thus, by Hempel's and (my) view, Lipton should modify his fudging argument in favor of the strong advantage thesis such that it only applies to certain types of hypotheses, that is, hypotheses that are difficult to assess in terms of complexity, hypotheses that are not or cannot be simplified into mathematically elegant formulas. It is not the time at which data is collected that gives rise to

different levels of support of a hypothesis, but rather the content and form of the hypothesis and the data.

## **VI. Concluding remarks**

My paper has hopefully illuminated some of the problems that arise in trying to justify the widely accepted notion that data that has been predicted provides stronger support for a hypothesis than data that has been accommodated, all else being equal. Many philosophers and scientists, like Lipton, point at the seemingly obvious risk of over-fitting accommodated data by fudging hypotheses to explain more of the data. However, explaining all of the data should be a goal of science, not something to be avoided. There is a risk of over-fitting accommodated data, but there is a similar risk of wrongly confirming or rejecting hypotheses using predicted data. And so it isn't clear why we should prefer predicted data to accommodated data. Finally, it seems like scientific practice should be objective. When scientists make hypotheses from their data, or use data to confirm or reject their hypotheses, it doesn't seem like the support of a hypothesis depends at all on whether the scientist is more or less likely to do bad science. So any argument that claims that accommodated data is worse than predicted data because accommodated data make scientists more likely to do bad science will not suffice because it will not be consistent with the objective nature of science.