Should We Act on Higher-Order Evidence?

Brian Weatherson

It's a platitude, or perhaps a cliché, that we should proportion our belief to the evidence. I think that platitude is basically correct. At least, I think it's more correct than many other philosophers do. But it is worth noting how oddly unstable a position it is. Consider a simple whodunit, where there are three suspects, the Administrator, the Butler and the Cook. Actually, the administrator did it. All the evidence, however, points to the butler. But the evidence the detectives have suggests that that very evidence points to the cook.[1]

You might worry already that the case is impossible. It's one thing for evidence to mislead about facts; it's another thing for the evidence to be self-misleading in this way. And indeed there are arguments, some vaguely Socratic, some vaguely Kantian, that the case is really impossible. But I'm going to work with the assumption that evidence could reasonably mislead one about any matter whatsoever, and ask how we should judge agents who find themselves in that situation. So imagine there are three detectives following the case.

- Detective A has a gut feeling that it is the Administrator, and has a high credence that she did it.
- Detective B follows the evidence, and has a high credence that the butler did it, although Detective B also, as a good follower of evidence, thinks it is very unlikely that the evidence supports her high credence.
- Detective C thinks that B is incoherent, and resolves the incoherence by thinking it most likely that C did it.

To judge that B is doing the best out of these three, you have to hold two views simultaneously. You have to think that just being right isn't the only consideration; a rational agent will be misled by misleading evidence. But when it comes to following the evidence, what matters is whether you actually follow it, not whether you think, or even have reasons to think, that you are following. I think all this is correct, but it's easy to feel otherwise.

In this talk I'm going to look at some recent philosophical work that is more sympathetic to Detective C's point of view. I'll start with an example, closely modelled on one offered by David Christensen, that is meant to reinforce that sympathy.

[1] This is the script for a talk I'm doing at the 2014 AAP. It isn't a paper. The citations, among other things, are not at the level of a paper. This material is taken from chapter 6 of my in progress manuscript *Normative Externalism*.

> Dr. D is a resident in a hospital. She sees a new patient, reads his chart, and forms the firm belief that he has disease X. This is right. Indeed, the patient has disease X, and that conclusion is well supported by the chart the doctor reads. Dr. D is about to order treatment for X when she is told several additional facts. She, i.e., Dr. D, has been on duty for a very long time. Doctors who have been on duty that long are often extremely over-confident in their diagnoses, even when they don't feel any internal doubt, as Dr. D does not. And if the patient has X, there is a reasonably cheap and reasonably quick test that can be run to confirm this before irreversibly committing to treatment.

It's easy to have the intuition that Dr. D should run the test, rather than start the treatment. If you don't share that intuition, increase the cost of providing the treatment for X. (Perhaps it involves amputation.) Or decrease the cost, in time, money and suffering, of waiting for the test results. Some days my intuitions still waver a little - after all, Dr. D had the right verdict for the right reasons - but I gather most people intuit that in this situation, Dr. D should run the tests, not start the treatment.

From there, it is easy to get an argument that we should not, in general, proportion our belief to the evidence. Here's one way to make it work.

1. If Dr. D proportions her beliefs to the evidence, she will have a very high credence that the patient has disease X.
2. If Dr. D has a very high credence that the patient has disease X, it will maximise expected utility to start treatment.
3. Dr. D should maximise expected utility.
4. So, if Dr. D proportions her beliefs to the evidence, then she will start treatment, not run more tests. (From 1–3)
5. Dr. D should run more tests, not start treatment.
6. So, Dr. D should not proportion her beliefs to the evidence. (From 4,5)

We could quibble over the details. The more extreme you make the treatment, in order to shore up the intuition behind 5, the less likely it will be that premise 2 is correct. So there needs to be some care taken to ensure that 2 and 5 are true together. And one could really doubt that premise 3 is true in this case. (It's hard to see how to square expected utility maximisation with the Hippocratic Oath.) More generally, it is certainly a live option to think that in cases where one is unsure what the evidence supports, expected utility maximisation is not a great idea. But I'm going to set all that aside, and instead argue against premise 1.

Even if the public evidence, in this case the chart, supports the conclusion that the patient has disease X, that doesn't mean that the totality of the evidence supports that

conclusion. There is also what we might call private evidence, namely the evidence the doctor gets by thinking through the case. More generally, I'm going to argue that many epistemologists have located the distinction between gathering evidence and processing evidence at nowhere near the right place.

I intend the reply I'm going to make to work not just against the particular argument I displayed above, but against a whole class of arguments based on examples like the one involving Dr. D. The recent literature features many, many such examples, often put forward with the intent of supporting something like the displayed argument. Interestingly, these examples all share some striking characteristics with the case of Dr. D.

1. They are all cases where the agent gets evidence that her cognitive skills may be less trustworthy than she antecedently thought.
2. They are all cases where the natural reaction to this higher-order evidence is to lower her confidence in the most salient proposition.
3. They are all cases where this natural reaction is to adopt a more cautious course of action.

We can imagine cases where the agent gets evidence about her evidence, or at least about her ability to process the public evidence, that have any subset of these characteristics. And when we do, the idea that cases like this can be used to reinforce Detective C's point of view becomes less plausible. Here's a simple variant that shares attribute 1, but not attributes 2 or 3, with Dr. D's case.

> Dr. E is a resident in a hospital. She sees a new patient, reads his chart, and forms the view that he probably has disease Y, but it is too early to be sure, for he might have disease Z. This is the attitude best supported by the chart; disease Z is rare, but not so rare as to just dismiss, and the recorded symptoms are completely consistent with disease Z. She is then informed of several things. She, i.e., Dr. E, has been awake for twelve hours, and is in her first week in a new ward. And doctors on their first week in a new ward have been found to be excessively cautious in their diagnoses, even when, like Dr. E, they think they have positive grounds for not settling on a firm diagnosis. Moreover, while there are tests that could be run to confirm the patient has disease Y before they start treatment, the patient will be suffering while the tests are run, and the treatment, if appropriate for the disease, will relieve that suffering.

I think it's reasonably clear that Dr. E should still order the further tests. If there is an alternative diagnosis that is compatible with the public evidence, and plausible given

the background knowledge of disease frequency, it would be wrong to rule it out just because other doctors similar to Dr. E are excessively cautious in their diagnosis. And it would certainly be a mistake to start a treatment regime because one discovered that others make a certain kind of error.

There is, then, an asymmetry between the two cases. In both cases, the doctor starts with a judgment based on public evidence, and forms a plan of action on the basis of this judgment. And in both cases, the judgment and the plan are rational responses to the public evidence. Then the doctor learns about a kind of error that is extremely prevalent amongst a class of which they are a representative member. In the first case, intuition says that should change at least the plan of action, and probably the underlying judgment. In the second, it says that the plan of action should be unchanged, and probably the underlying judgment should be as well. This asymmetry is something that could do with explaining. To anticipate, I propose to explain it using the notion of private evidence.

In the interests of finishing on time, I'm not going to defend this, but I think the kind of view I attributed above to Detective C doesn't have the resources to explain the asymmetry. And this is, I think, bad news for so-called conciliationist views of disagreement, and for similar views about what our detectives and doctors should do. But I won't argue for this. Instead, I'll put forward a very different explanation of the asymmetry, and leave questions of whether it is the best possible explanation for another day.

I'm going to set up my explanation by thinking through some examples involving mathematics. Let's start with a rather extreme example. A young mathematics student, Frida, is told about Fermat's Last Theorem. Frida knows enough to understand exponentiation. So what she knows about basic arithmetic entails that Fermat's Last Theorem is true, though most people in her position would not believe this. (Indeed, it entails both that the Theorem is true, and that the usual statements of the theorem express truths.) Frida is not told, and does not know, about the long history of Fermat's Last Theorem, and about how it puzzled mathematicians for centuries.

Frida spontaneously forms the belief that Fermat's Last Theorem is true. This isn't because Frida is usually prone to making mathematical guesses – this is the only time that she forms a mathematical belief about such a theorem in such a spontaneous way. But nor is it because she is a mathematical savant; she can't articulate anything like a proof of Fermat's Last Theorem, or hint at what a proof of it would look like. This all suggests that she does not know the theorem is true. To be sure, the inability to articulate a proof doesn't entail that one does not know it is true. I know Fermat's Last Theorem is true, but could not prove it. But given Frida's situation, it does seem that she fails to know the Theorem is true.

Frida's lack of knowledge here is something that needs explaining. I think that the best explanation for it is that she lacks evidence. What other explanations are available?

Could it be that Frida is unreliable? It doesn't seem so; by hypothesis she is perfectly reliable, at least in the actual world. Could it be that Frida's method is unreliable? Again, it doesn't seem so. The method of spontaneously forming mathematical beliefs about mathematical propositions is one that we use every day, and usually reliable. It's true that if we restricted the scope of the methods to 'complex' propositions, it would be true that Frida's method is unreliable. But I don't think there's any way of giving a sense in which Fermat's Last Theorem is complex without appeal to some independent epistemological notion. It isn't, after all, complex to state. What is complex is proving it. The point here is that the only way to draw a reference class of exactly the right breadth to make Frida's belief-forming method unreliable is to use other epistemological ideas, like the idea that Fermat's Last Theorem is hard to prove. That suggests it is the other epistemological ideas that explain Frida's lack of knowledge, not the unreliability of his method relative to a somewhat gerrymandered reference class. And note that even if we get the reference class right, we have to somehow make it the case that the fact that Frida is herself quite reliable when spontaneously forming mathematical beliefs, even about complex cases, irrelevant. I don't claim to have sealed the case, but I think it is going to be hard to explain Frida's ignorance in terms of unreliability.

So I think the best explanation of Frida's ignorance is that she lacks sufficient evidence to know that Fermat's Last Theorem is true. When one works through a mathematical, or a logical, or a philosophical, problem, one gets evidence about the correct answer to that problem. But Frida lacks that evidence. And she doesn't have any other form of evidence, like the testimonial evidence most of us have, that would make up for the lack of mathematical evidence. So thinking through a mathematical problem is not just a matter of processing evidence, it is a way of acquiring new mathematical evidence. It is because Frida lacks just that evidence that she lacks knowledge.

This way of thinking about mathematics is hardly radical. It is a commonplace in mathematics that one can get evidence for or against mathematical propositions. Here's one nice example. Sanjoy Mahajan (2010) describes a lot of heuristics that can be used to quickly refute various mathematical hypotheses. The heuristics involve, for example, checking whether the 'dimensions' of a proposed identity are correct, checking limit cases, and that sort of thing. So consider the hypothesis that the area of an ellipse is $\pi a b$, where a is the distance from the centre to the nearest point on the ellipse, and b is the distance from the centre to the furthest point. After going through a number of other proposals and showing how they can be refuted by some easy to apply tests, Mahajan says this about the proposal that the area is $\pi a b$.

> This candidate passes all three tests. …With every test that a candidate

> passes, confidence in it increases. So you can be confident in this candidate. And indeed it is correct. (Mahajan, 2010, 21)

In familiar terminology, seeing that the hypothesis that the area is $\pi a\, b$ passes Mahajan's tests is a way of gathering evidence that the hypothesis is true.

So I think we might need extra evidence to properly infer a conclusion from some premises, even when the premises entail the conclusion. This is really just a version of Gilbert Harman's dictum that inference is not implication. (Harman, 1986) But it might be worried that it runs into the regresses described by Lewis Carroll (1895). It certainly would be bad to say that to infer q from p and $p \rightarrow q$, the agent needs to know $(p \wedge (p \rightarrow q)) \rightarrow q$. That way lies regress, and perhaps madness. But that's not what is being claimed here. Rather, the claim is that for non-obvious entailments, the agent must know that the entailment obtains, or at least know the corresponding material implication, in order to use it in reasoning. It's consistent with that that we don't need any extra evidence to make simple inferential moves, either in deductive or inductive reasoning. And that's all we need to say to stop the regress.

With this view of evidence in mind, let's look again at the case of Dr. D. One crucial detail of the case that is not filled in. I said anyone in Dr. D's position would judge it was disease X. We now need to ask why this is so. Here are three options.

1. Any reasonable person, even without any training in medicine, would diagnose X given the public evidence.
2. Option 1 isn't true, but the correctness of the diagnosis is immediately entailed by an easy observation, plus a piece of knowledge Dr. D has.
3. Neither of the first two options is true, and Dr. D needs to either use developed skills at observation, or put together many pieces of knowledge in a complex inference.

I'll start with option 3, which I think is the most normal situation. That's because in medicine, there are few universal generalisations of the form *Everyone with symptoms S has disease X*, at least where those symptoms can be easily observed. Rather, there are true generics. So making a good diagnosis involves relying on situation-specific knowledge of propositions like *There are no other facts that I need to learn before making a diagnosis*, and *There are no other plausible explanations of the data*. If these claims are true, but not known, then Dr. D is in a situation like Frida's. The public data supports the conclusion she draws, but unless she goes through the appropriate steps to get from the public data to the diagnosis, she doesn't have a justified belief in the diagnosis. That's to say, she needs both the facts about the patient's appearance, test results, and so on, plus the facts she learned in her medical training, plus these extra 'negative' facts, plus facts about what entails or supports what, in order to form a justified belief in the

correct diagnosis. And this is still, I suspect, a gross simplification of any realistic case. But the extra complications will only strengthen the point I want to make.

Once we see what Dr. D needs to know to make the diagnosis, it is easy to see how the higher-order evdience could be relevant. If the higher-order evidence is accurate, then the resident's beliefs in these further facts needed to make the diagnosis are highly unreliable. As such, they don't amount to knowledge. And so they aren't part of her evidence. So she doesn't have sufficient evidence to make the diagnosis. Note that in the simplest version of Dr. D's case, all that will be true before the resident is reminded of how long she has been awake; she'll simply never have a justified belief. If the higher-order evidence is misleading, then Dr. D will be forming these extra beliefs by a reliable method, but it will be a method she has good reason to believe is unreliable. And one can't get knowledge by using a method one has good reason to believe is unreliable. So Dr. D still doesn't know the relevant background facts, so still doesn't have enough evidence to make the diagnosis. Note that in this case, she may have originally known that the diagnosis was correct, but then lost the knowledge once she got reason to believe that her methods for forming beliefs in relevant background facts was unreliable.

I've tried to motivate this story by going via the example of Frida, to note the importance of thinking in evidence collection. But we could also motivate it by thinking about Dr. E. We needed to explain why learning that she may be excessively cautious does not motivate action. And now we have the basis of an explanation.

In order to form a belief that her patient has disease Y, Dr. E would need to have, as part of her evidence, that there are no reasonable alternative explanations of the data about the patient. And that means she needs to know that is true. But she can't know it is true, because it simply isn't true. It might be that she could be misled into reasonably believing that there are no reasonable alternative explanations, and that she is being unreasonable in being cautious. But in order to have sufficient evidence to believe the patient has disease X, she would need to have in her evidence something that is not true. And while higher-order evidence that one is excessively cautious can make a falsehood reasonable to believe, it can't make it into knowledge. And without making it into knowledge, it can't make it into evidence. And that means Dr. E can't get the evidence she needs to diagnose the patient with Y.

In short, thinking about the importance of private evidence, and about how learning about one's own unreliability can destroy private evidence, gives us a nice explanation of the asymmetry between the doctors. Explanations of Dr. D's case that give up on evidentialism do not have as natural an explanation of the asymmetry. This tells against the use of the Dr. D case to undermine evidentialism.

There is one loose end to tie up. It is a consequence of my theory of these cases that when the public evidence directly supports a conclusion, an agent who draws that

conclusion need not change their views in the face of higher-order evidence. If someone knows that they owe £2 + £2, then they can know that they owe £4, even if they have evidence that they have been exposed to a drug that makes people poor at arithmetic. That's because they don't need to rely on any extra evidence to get from the premise that they owe £2 + £2 to the conclusion that they owe £4. In cases of direct evidential support like this, they get justification as long as they get the conclusion right. (This is a way in which the theory of the cases I'm giving is externalist; for immediate inferences, correctness is the only requirement.)

If Dr. D's case is as easy as this problem, then the story I'm telling doesn't explain why she should change her plan of action on learning how long she has been awake. That's true whether the puzzle is as easy as what is £2 + £2 for anyone, or for anyone with Dr. D's knowledge. But I think that's not a problem for three reasons.

1. The intuitions about the case are not as clear as in the standard version of the case.
2. In related cases involving peer disagreement, even those philosophers who have said that higher-order evidence should change one's beliefs put in exception clauses for problems this simple.
3. The main argument that Dr. D can't simply use her knowledge in this case is that it leads to a kind of circular reasoning. But it turns out to be impossible to formulate a plausible anti-circularity principle that Dr. D would thereby violate.

Sadly, there isn't time to argue for point 3. So I hope points 1 and 2 are enough to convince you!

References

Carroll, Lewis. 1895. "What the Tortoise Said to Achilles." *Mind* 4:278–280.

Harman, Gilbert. 1986. *Change in View*. Cambridge, MA: Bradford.

Mahajan, Sanjoy. 2010. *Street-Fighting Mathematics: The Art of Educated Guessing and Opportunistic Problem Solving*. Cambridge, MA: MIT Press, second edition.