I Can’t Believe I’m Stupid[1]

Andy Egan and Adam Elga

Australian National University/University of Michigan and Princeton University

It is bad news to find out that one's cognitive or perceptual faculties are defective. For one thing, it’s news that something bad is happening – nobody wants to have defective cognitive or perceptual faculties. For another thing, it can be hard to see what to *do* with such news. It’s not always transparent how we ought to revise our beliefs in light of evidence that our mechanisms for forming beliefs (and for revising them in the light of new evidence) are defective.

We have two goals in this paper: First, we’ll highlight some important distinctions between different varieties of this sort of bad news. Most importantly, we want to emphasize the following distinction: On the one hand, there is news that a faculty is *unreliable*--that it doesn't track the truth particularly well. On the other hand, there is news that a faculty is *anti-reliable*--that it tends to go positively wrong. These two sorts of news call for extremely different responses. Our second goal is to provide rigorous accounts of these responses.

*

We begin with an easy case: ordinary, garden variety news of unreliability.

Sadly, we don’t have to look far for examples. Take, for instance, the deterioration of memory with age. As you increasingly call students by the wrong names, you begin to think that your memory for names is not what it once was. How should this news of unreliability affect the strength of your beliefs about who has what names? Clearly it would be irresponsible to retain the same degree of confidence that you had before you got the bad news. On the other hand, it would be overreacting to become completely agnostic about which names people bear. What is in order is a modest reduction in confidence.

For instance, across the room is a student—--you seem to remember that her name is Sarah. Your decreased trust in your memory should slightly reduce your confidence that her name is Sarah.

But in addition to reducing the strength of your beliefs about who has what names, the news should also reduce the resiliency of those beliefs (Skyrms 1977). In particular, the news should make it easier for additional evidence to further reduce your confidence that the student is named Sarah. To bring this out, suppose that from across the room, you hear a third party call the mystery student "Kate". Back when you thought your memory was superb, hearing this would have only slightly reduced your confidence that the student's name was Sarah. (In those days, you would have thought it likely that you'd misheard the word "Kate", or that the third party had made a mistake.) But now that you count your memory as less reliable, hearing someone refer to the mystery student as "Kate" should significantly reduce your confidence that her name is Sarah. You should think it fairly likely that you misremembered. This illustrates the way in which news of unreliability should reduce the resiliency—and not just the strength—of your beliefs about names.

How much reduction in strength and resiliency is called for? This of course depends on how compelling the news of unreliability is, and on the strength of the competing source of information. It is worth working through a simple example to see how things go.

Suppose that you're certain that the student in question is either named Sarah or Kate. Think of your memory as a channel of information with 99% reliability: it had a 99% chance of making and sustaining a correct impression of the student's name. (We're starting with the case in which you count your memory as being superb.) And think of the words that you overhear across the room as an independent channel of information, with 95% reliability.

Initially, your memory indicates that the student's name is Sarah, and so you believe this to degree .99. (Here it is assumed that, independent of your memory impressions, you count each name as equally likely.) But when you hear someone call the student "Kate", you become less confident that she is named Sarah. A quick Bayesian calculation[2] shows that your new level of confidence is .84.

So: in the "superb memory" condition, you start out very confident that the student is named Sarah (.99), and this confidence is reduced only modestly (to .84) when you overhear "Kate". How would the calculation have gone if you had counted your memory as less than 99% reliable? Suppose, for example, that you had counted your memory as being merely 90% reliable. In that case, your initial degree of belief that the student was named Sarah would have been .9 — slightly lower than the corresponding degree of belief in the "superb memory" condition. Now let us consider how resilient that .9 would have been. That is, let us check how much your confidence would have been reduced upon overhearing someone call the student "Kate". Answer:[3] your new level of confidence would have been .32.

This low value of .32 brings out a striking contrast. In the "superb memory" condition, your level of confidence that the student was named Sarah was fairly resilient. But in the "just OK memory" condition, that level of confidence is not at all resilient: overhearing "Kate" in this condition massively reduces your confidence that the student is named Sarah. See Figure 1.

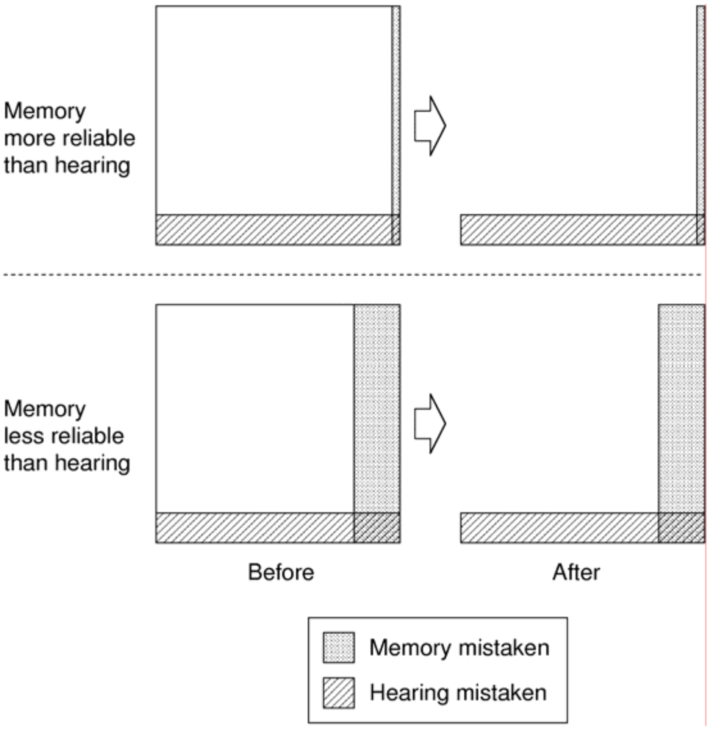

Figure 1: How reduced trust in your memory makes your memory-based beliefs much less resilient. You are attempting to determine whether a particular student is named Sarah or Kate. Initially, you seem to remember that the student is named Sarah. Each row of the figure shows how your confidence in this memory is reduced when you seem to overhear the student addressed as “Kate”. In the top row, you are initially extremely confident in your memory—as is reflected by the extreme narrowness of the shaded region of the upper-left square (regions correspond to propositions, and each propositions has an area proportional to its probability).

Seeming to hear the student addressed as “Kate” indicates that either your memory or your hearing is mistaken—which corresponds to ruling out the blank region, in which both your memory and your hearing are correct. As a result, your confidence that your memory is mistaken only increases slightly (since the shaded region only occupies a small proportion of the remaining area). In contrast, if you had started out with less initial confidence in your memory (bottom row), seeming to overhear “Kate” would have drastically increased your confidence that your memory was mistaken, since erasing the blank region would leave more than half of the remaining area shaded.

The above is merely a toy example, but the lesson generalizes:

When one initially counts a channel of information as extremely reliable, a small reduction in that reliability should (a) slightly reduce your confidence in beliefs deriving from that channel, but (b) massively reduce the resiliency of those beliefs.

*

The above is what happens in general when we get evidence that some source of information is unreliable – it wasn't important that the source of information was one of our own cognitive mechanisms. Exactly the same thing happens when our confidence in more external sources of information – a newspaper, an informant, a doctrinal text – is undermined

The case of the doctrinal text is particularly interesting: one piece of practical advice that emerges from the above discussion is that, when some faith (either religious or secular) is based on a particular revered and authoritative text, a good way to undermine that faith is to first convince the faithful that the text isn't so authoritative after all, rather than simply arguing against particular points of doctrine. So long as the flock thinks that the text is the product of divine inspiration (or of superhuman genius), they will likely count it as extremely reliable, which will make them relatively immune to corruption by other sources of evidence. But even small reductions in how much they trust the text will make the flock much more easily convinced that particular bits of doctrine are wrong. On the other side of the coin, it may be that the best way to protect such a text-based faith is not to defend the points of doctrine piecemeal, but to defend the infallibility (or at least the near-infallibility) of the text.

*

The memory example above was particularly clean. In it, the news you received concerned the unreliability of only a single faculty (memory for names). Furthermore, your ability to get and respond to the news did not rely on the faculty in question.

Because the news was confined to a single faculty, it was easy to "bracket off" the outputs of that faculty, and to thereby see which of your beliefs deserved reduced strength and resiliency as a result of the news. The same would go for news of the unreliability of any perceptual system or reasoning process in a well-delineated domain. For example, one might find that one tends to misread the orders of digits in phone numbers. Or one might find that one is particularly unreliable at answering ethical questions when one is hungry. In each such case, the thing to do is to treat the outputs of the faculty in question with the same caution one might treat the display of an unreliable wristwatch.

Things aren't always so simple. Consider, for example, a case in which news of unreliability arrives by way of the very faculty whose reliability is called into question:

> Your trusted doctor delivers an unpleasant shock. "I am afraid that you have developed a poor memory for conversations," she says. "Overnight, your memories of conversations had on the previous day often get distorted." The next day, you wake up and recall the bad news. But you aren't sure how much to trust this memory. As you trust it more, you begin to think that you have a bad memory—and hence that it doesn't deserve your trust. On the other hand, as you doubt your memory, you undermine your reason for doing so. For the

> remembered conversation is your only reason for thinking that you have a poor memory.[4]

Neither resting place (trusting your memory, or failing to trust it) seems stable. Yet there should be some reasonable way for you to react. What is it?

The answer is that you should partially trust your memory. As an example, let us fill in some details about your prior beliefs. Suppose that you were antecedently rather confident (90%) in the reliability of your memory. Suppose that conditional on your memory being reliable, you counted it as very unlikely (1%) that you would remember your doctor reporting that your memory was unreliable. And suppose that conditional on your memory being unreliable, you thought it quite a bit more likely that you'd remember your doctor reporting that your memory was unreliable (20%). Then an easy calculation[5] shows that when you wake up the day after your doctor's visit, your confidence that your memory is reliable should be approximately .31.

The resulting state of mind is one in which you have significant doubts about the reliability of your memory. But this is not because you think that you have a reliable memory of yesterday's conversation. Rather, it is because your memory of the conversation is simultaneous evidence that (1) your memory is unreliable and (2) it has happened to get things right in this particular case. (Note that since there is a tension between (1) and (2), the reduction in trust in your memory is not as dramatic as it would have been if you had possessed, for example, a written record of your conversation with the doctor.)

The above case is analogous to the case of a self-diagnosing machine, which reports periodically on its own status. When the machine outputs “I am broken”, that might be evidence that a generally unreliable self-diagnostic process has gone right on this particular occasion, so that the machine is faithfully (though not reliably) transmitting the news of its own unreliability. Alternatively, the output might have the same kind of evidential force as the output “I am a fish stick”: it might be evidence that the machine is broken simply because working machines are unlikely to produce such output. Either way, we have an example of a mechanism delivering news of its own unreliability, without that news completely undermining itself.

The above is, again, only a toy example. But again, the lesson generalizes:

News of unreliability can come by way of the very faculty whose reliability is called into question. The news need not completely undermine itself, since the reasonable response can be to become confident that the faculty is unreliable, but has happened to deliver correct news in this instance.

*

Now let us turn to anti-reliability.

In the simplest case, news of anti-reliability is easy to accommodate. Consider a compass, for example. When the compass is slightly unreliable, the direction it points tends to be close to North. When the compass is completely unreliable, the direction it points provides no indication at all of which direction is North. And when the compass is anti-reliable, it tends to point *away* from North. Upon finding out that one's compass is anti-reliable, one should recalibrate, by treating it as an indicator of which direction is *South*.

Similarly, one might learn that a perceptual or cognitive faculty is anti-reliable, in the sense that it delivers systematically mistaken or distorted outputs. For example, one might find that when judging whether a poker opponent is bluffing, one's initial instinct tends to be wrong. Here, too, one should recalibrate. For example, one should treat the initial hunch “Liz is bluffing” as an indication that Liz is not bluffing.

Other cases are trickier.

One of the authors of this paper has horrible navigational instincts. When this author—call him "AE"—has to make a close judgment call as to which of two roads to take, he tends to take the wrong road. If it were just AE's first instincts that were mistaken, this would be no handicap. Approaching an intersection, AE would simply check which way he is initially inclined to go, and then go the opposite way. Unfortunately, it is not merely AE's first instincts that go wrong: it is his all-things-considered judgments. As a result, his worse-than-chance navigational performance persists, despite his full awareness of it. For example, he tends to take the wrong road, even when he second-guesses himself by choosing against his initial inclinations.

Now: AE faces an unfamiliar intersection. What should he believe about which turn is correct, given the anti-reliability of his all-things-considered judgments? Answer: AE should suspend judgment. For that is the only stable state of belief available to him, since any other state undermines itself. For example, if AE were at all confident that he should turn left, that confidence would itself be evidence that he should not turn left. In other words, AE should realize that, were he to form strong navigational opinions, those opinions would tend to be mistaken. Realizing this, he should refrain from forming strong navigational opinions (and should outsource his navigational decision-making to someone else whenever possible).[6]

Moral:

> When one becomes convinced that one's all-things-considered judgments in a domain are produced by an anti-reliable process, one should suspend judgment in that domain.

*

When AE faces an intersection, what forces AE into suspending judgment is the following: his decisive states of belief undermine themselves. For it would be unreasonable for him to both make a navigational judgment and to think that such judgments tend to go wrong. In other words, it is unreasonable for AE to count himself as an *anti-expert*—someone whose state of belief on a given subject matter is quite far from the truth (Sorensen 1988, 392). And this is no mere special case: *It is never reasonable to count oneself as an anti-expert*.[7] It follows that there are limits on the level of misleadingness one can reasonably ascribe to one's own faculties, no matter *what* evidence one gets. Let us sharpen up and defend these claims.

Start with anti-expertise. It is never rational to count oneself as an anti-expert because doing so must involve either incoherence or poor access to one's own beliefs. And rationality requires coherence and decent access to one's own beliefs.

The latter claim—that rationality requires coherence and decent access to one's own beliefs—we shall simply take for granted.[8] Our interest here is in defending the former claim: that counting oneself as an anti-expert requires either incoherence or poor access to one's own beliefs.

Start with the simplest case: the claim that one is an anti-expert with respect to a single proposition. For example, consider the claim that one is mistaken about whether it is raining:

> (M) Either it is raining and I believe that it is not raining, or else it is not raining and I believe that it is raining.

No consistent agent with perfect access to her own beliefs believes M. For if such an agent believes that it is raining, she also believes that she so believes—and together these two beliefs are inconsistent with M. The same goes if she believes that it is not raining. The only other possibility is that she suspends judgment on whether it is raining, in which case she believes that she suspends judgment—which is also inconsistent with M.

The bottom line is that no consistent agent with perfect access to her own beliefs believes that she is an anti-expert with respect to a given single proposition. This is familiar news, since the relevant claim of anti-expertise is a close cousin of the famously Moore-paradoxical claim:

> It is raining but I don't believe that it is. (Sorensen 1988, 15)

What is less obvious, however, is that the same Moore-paradoxicality infects claims of anti-expertise in more general settings. For example, when an entire subject-matter (not just a single proposition) is in play, the claim that one is an anti-expert with respect to that subject-matter is also Moore-paradoxical. And this result is no mere artifact of treating belief as an all-or-nothing matter: it continues to hold when degrees of belief are taken into account. Furthermore, we can be precise about the maximum degree to which one can reasonably believe that one is an anti-expert. Finally, none of these conclusions require the assumption of *perfect* access to one's own state of belief: they hold even under the assumption of *good* access to one's own state of belief.

Showing all of this will take some doing.

*

To be an anti-expert on a subject matter is to have beliefs on the subject matter that are inaccurate—quite far from the truth. But what exactly is it for a state of belief to be inaccurate with respect to a subject matter?

Here is one simple answer. Restrict attention to a finite list of propositions, chosen to represent a subject matter. Let us say that an agent is an anti-expert with respect to those propositions if (1) the agent is *confident in* at least one of them, in the sense that his degree of belief in it is at least 90%; and (2) at least half of the propositions (on the list) that the agent is confident in are false. For example, for AE to be an anti-expert about geography—as represented by an appropriate list of geographical propositions—is for him to be confident in at least one of those propositions, and at least half of the ones he is confident in to be false.

It turns out that the claim "I am an anti-expert about geography" is *unbelievable*, in the following sense. No coherent agent who is correct about her own beliefs believes it to degree greater than 20%.[9] For the details, we refer interested readers to Figure 2 and Appendix A, but the guiding idea can be illustrated in the special case of a subject matter that consists of exactly two propositions. Suppose that such an agent has 90% confidence in each of the two propositions. Then she must be at least 80% confident in their conjunction (for the same reason that two carpets, each of which covers 90% of the floor space of a single room, must overlap on at least 80% of the space). So she must be at least 80% confident that she is correct in both of her confident beliefs on the subject matter—and hence can be at most 20% confident that she is an anti-expert.[10]

The bottom line: no coherent agent who is correct about her own beliefs has degree of belief higher than 20% that she is an anti-expert.

Now, the numerical details of the above result depend on spelling out anti-expertise in a particular way. And the way in which we've spelled it out is admittedly not very sophisticated. But fancier measures of inaccuracy change only the details: using them leads to results of the same qualitative nature. For example, one might gauge the accuracy of a state of belief by its Brier Score, a measure used to assess probabilistic weather predictions (Brier 1950, as cited in Joyce 1998). The Brier score measures inaccuracy on a scale of 0 (not at all inaccurate) to 1 (completely inaccurate).[11] It can be shown that every coherent agent who is correct about her own degrees of belief assigns low probability to the hypothesis that she has a high Brier score. (For instance, such an agent always has probability less than .3 that her Brier score is greater than .6. See Figure 2 and Appendix B.)

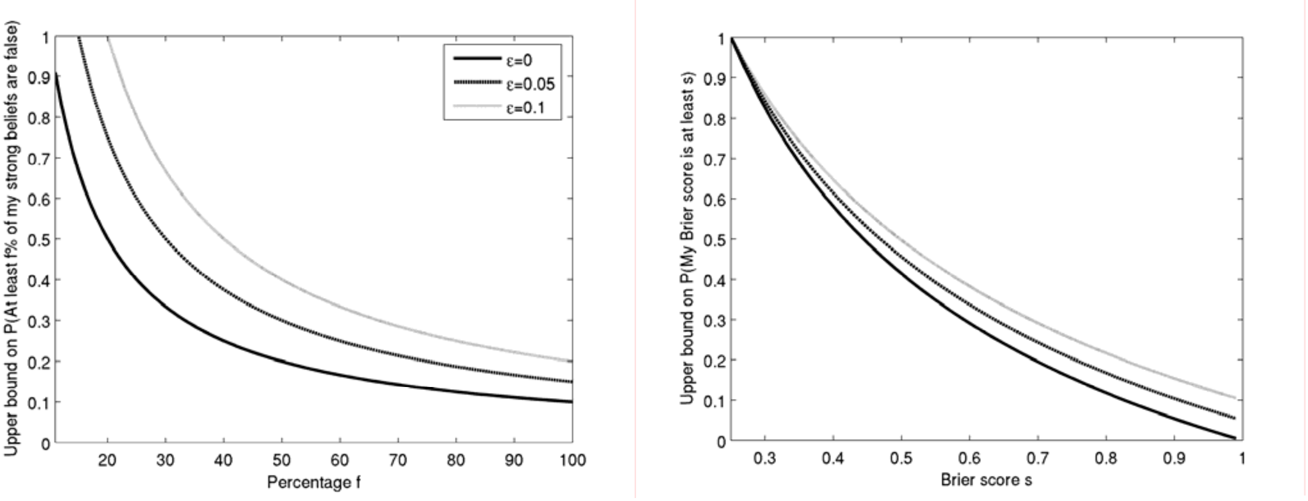

Figure 2: Upper bounds on probability of anti-expertise. Left: For each percentage *f*, graphed is an upper bound on the probability one can reasonably have that at least *f%* of one's confident (*≥90%*) beliefs are false. Right: For each score *s*, graphed is an upper bound on the probability one can reasonably have that one's Brier score is at least *s*.

In each case, it is assumed that reasonable agents are mistaken about their own degrees of belief by no more than ε, and graphs are shown for ε=0 (solid line), .05 (dashed line), and .1 (dotted line).

The above results concern agents who are *perfectly* correct about their own degrees of belief. Some consider this assumption too strong, even for ideally rational agents (Williamson 2000), (Williamson forthcoming). So it is worth noting that the results are perturbed only slightly if we modestly relax the assumption of perfect introspective access. For example, we might restrict attention to coherent agents who are *nearly* correct about their own degrees of belief. It remains true that every such agent assigns low probability to the proposition that she is an anti-expert. (For the details, see appendices A and B.)

The bottom line:

Given a fixed subject matter, any coherent agent who has decent access to her own beliefs assigns a low probability to the claim that she is an anti-expert about that subject matter. It follows that it can never be rational to count oneself as an anti-expert about a given subject matter.

*

It is unreasonable to count oneself as an anti-expert. In contrast, it can be perfectly reasonable to count someone else as an anti-expert. This contrast can seem puzzling. Why treat evidence about oneself so differently than evidence about others?

For example, consider Professor X. Readers of Professor X's books on geology note that her geological opinions are far from the truth. They also note that each book she publishes is as inaccurate as its predecessors, even though in each case Professor X acknowledges her past inaccuracy and tries to correct matters. Such readers may reasonably count Professor X as an incorrigible anti-expert about geology.

Now suppose that Professor X receives an anonymized version of her own track record: she is told that a certain unnamed scholar has produced a number of books, that the books have been persistently inaccurate, and so on. Couldn't Professor X reasonably conclude that the scholar in question is an anti-expert? And couldn't that judgment reasonably persist, even if she were told that the scholar in question was Professor X herself?

The alternative seems to be that rationality requires an illegitimate sort of chauvinism—that it requires Professor X to say, "When I thought that the unnamed scholar was someone else, then I had excellent evidence that this unfortunate soul was an incorrigible anti-expert. But not if the unnamed scholar was *me*!" That looks like pure arrogance. Yet it seems to be the response we must recommend, if we are to insist that one can never reasonably count oneself as an anti-expert. What has gone wrong?

What has gone wrong is that Professor X's response has been misrepresented. When Professor X finds out that the hapless scholar is her, she should become convinced that she has been an anti-expert. But she should do something about it: she should change her opinions about geology in such a way that she is no longer an anti-expert. For the news of her past anti-expertise is evidence that her present geological views are mistaken.

In contrast, if Professor X finds out that someone *else* has been an anti-expert about geology, that need not count at all against Professor X's present geological views. For example, it may be perfectly reasonable for Professor X to think: "Dr. Z is an incorrigible anti-expert about geology. His present geological opinions are quite far from the truth." This attitude may be reasonable because Professor X's geological opinions may be quite different from those of Dr. Z. Learning that Dr. Z's geological opinions are largely false will undermine Professor X's beliefs about geology only to the extent that Professor X, prior to receiving this news, agreed with Dr. Z.

A comparison with a simpler case is helpful here. Suppose that you are informed, "Before you received this news, you were mistaken about whether it is raining". In response, you ought to change your mind about whether it is raining. In contrast, if you are informed, "Before you received this news, *Fred* was mistaken about whether it is raining", you need be under no pressure at all to change your beliefs about the rain.

When we find out that the anonymous anti-expert is *us*, we will (if we're rational) revise the anti-expert's opinions – we will distance our new opinions from our old, discredited ones. We will then cease to be anti-experts. When we find out that the anonymous anti-expert is someone else, the situation is different. When I discover that Fred's current beliefs about some subject matter are systematically mistaken, my revision will distance *my* new opinions from Fred's old, discredited ones. But *my* revision won't put any distance between *Fred's* new opinions and his old ones. In general, it will be reasonable for me to expect that Fred has retained his discredited beliefs. And so it will be reasonable for me to believe that Fred is an anti-expert. There is no asymmetry in the degree to which the subject's current views are undermined by evidence of anti-expertise. In both cases, the recipient of the news should distance their beliefs from those that the subject held at the time the news was received. The asymmetry springs from the fact that, in the first-person case, doing this distancing will cure the subject's anti-expertise, while in the third-person case it will not.

How exactly should Professor X change her geological views in the light of her dismal record? That depends on her assessment of her own abilities to recalibrate in such situations. If she trusts her ability to recalibrate, then she should do so. After doing so, she will have a state of geological opinion that she can stand behind. She will judge that she *was* an anti-expert, but that she is one no longer.

It might be that Professor X does not trust her ability to recalibrate. For example, the record might show that her past attempts to do so have been unsuccessful, and that another attempt is unlikely to do better. In that case she is in the same situation with respect to geology that AE was with respect to navigation. She should think that any confident opinions she forms about geology are liable to be mistaken. So she should suspend judgment about geological matters.

Suspending judgment guarantees that Professor X is no longer an anti-expert, since anti-expertise requires not just ignorance, but error. For example, one can only be an anti-expert about a subject matter if one holds at least one belief in the subject matter with confidence at least 90%. Similarly, one can only have a large Brier score by having many probabilities that are close to 0 or 1. More generally, one can only have a state of opinion that is extremely far from the truth by being confident in many falsehoods.

If Professor X suspends judgment, she may reasonably believe that she *was* an anti-expert (back when she had strong geological opinions). She may reasonably believe that she *would* be an anti-expert (if she were to form such opinions). She may even reasonably believe that she *will* be an anti-expert in the future (because she expects to be unable to resist forming such opinions). But this all falls short of believing that she is *presently* an anti-expert. And that she may not reasonably do.[12]

The bottom line is that one *should* treat news of someone else's past anti-expertise in a different way than one treats the corresponding news about oneself. But this is not due to chauvinism or arrogance. Rather, it is because news of one's own anti-expertise on a subject matter is inevitably evidence against one's current views on the subject matter. And no such inevitability holds in the case of news about someone else.

*

On one natural way of understanding what it takes to be stupid, stupidity is productively contrasted with ignorance. The ignorant person lacks information, and so lacks strong opinions. The stupid person has strongly held false opinions, gotten as a result of lousy reasoning. In one sense, ignorance and stupidity are possibilities for me: other people, similar to me in relevant respects, can be both ignorant and stupid. I don’t take my cognitive faculties to be constructed in some special way that precludes either ignorance or stupidity. But there is also a sense in which ignorance, but not stupidity, is a possibility for me: I can reasonably believe that I’m ignorant, but I can’t believe I'm stupid.

Appendix A

The main text introduces the notion of an agent being an anti-expert about a subject matter, in the sense that at least half of the agent's confident beliefs about the subject matter are false. Here we introduce the more general notion of an agent being f-wrong about a subject matter (where f is some fixed fraction, not necessarily equal to one half), and give a bound for how confident a reasonable agent can be that he is f-wrong.

The bound derives from the following lemma.

Carpet lemma: Suppose that you are given n carpets, each of which occupies at least fraction x of a room with unit area. Your job is to lay all of the carpets in order to minimize A, the area that is covered by more than m layers of carpeting, where $m<n$. You may cut and overlap carpets if necessary, but no two pieces of the same carpet may overlap. Claim: You can do no better than $A = \mathrm{MAX}\{0,(nx-m)/(n-m)\}$.

Proof: Consider the most favorable case, in which each carpet occupies exactly fraction x of the floor space. Then the total area of carpet to be placed is nx. Your most efficient carpet-laying strategy is as follows. First, fill the entire room uniformly to depth m. If $nx \leq m$, this task will exhaust all of the carpet, and you will have completed your task with $A=0$. Otherwise, you will have laid area m of carpet, which leaves area $nx-m$ of extra carpet remaining. You wish to minimize the area that this additional carpet covers, which means piling it up as high as possible over a single region. (See Figure 3 for an example.) But piling higher than n total layers is disallowed, since there are n carpets and no two pieces of the same carpet are allowed to overlap. So you must stash the extra carpet in a pile at most $n-m$ layers high. So the extra carpet must cover an area of at least (extra carpet left/additional layers allowed), which equals $(nx-m)/(n-m)$.

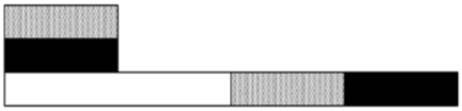

Figure 3: How to lay 3 carpets of area 1/2 in order to minimize the area covered by more than one carpet.

Let us use this lemma to derive our bound.

First some terminology and assumptions: Restrict attention to a finite set D of propositions, chosen to represent a subject matter. Let us say that an agent is *f-wrong* with respect to D if (1) the agent is *confident in* at least one member of D, in the sense that his degree of belief in it is at least x; and (2) more than fraction f of the propositions in D that the agent is confident in are false.

Assume that the agent in question has probability function P, and that he has good access to his own beliefs, in the following sense. For any proposition in D, the agent's actual degree of belief in that proposition is within ε of the degree of belief he thinks he has in it. In other words, the agent is certain that he has probability function Q, where for any X_i in S, $|Q(X_i)-P(X_i)| \leq \varepsilon$.

Claim: The agent's probability that he is f-wrong about D is no greater than $(1-x+\varepsilon)/f$.

Proof: Let S be the set of propositions in the subject matter that the agent believes she is confident in. If S is empty, then the agent has probability zero that he is f-wrong, and we are done. Otherwise, let n be the cardinality of S. Then the agent's probability that he is f-wrong is

P(I am f-wrong about S) = P(fraction of propositions in S that are false > f)

= 1-P(fraction of propositions in S that are false <= f)

= 1-P(fraction of propositions in S that are true > 1-f)

= 1-P(number of propositions in S that are true > m),

where $m=(1-f)n$. We will get an upper bound on P(I am f-wrong about S) by getting a lower bound on P(number of propositions in S that are true > m). Without loss of generality, assume that m is an integer.

Think of logical space as a room with unit area, and think of each member of S as a carpet. In the analogy, minimizing the probability that more than m of the members of S are true corresponds to minimizing the area that is covered by more than m layers of carpet, under the conditions of the carpet lemma. Since the agent thinks she has at least probability x in each member of S, she really has probability of at least $x-\varepsilon$ in each member of S (by the assumption of good access). So we can apply the carpet lemma, assuming that each carpet has an area of at least $x-\varepsilon$. The result is that P(number of true propositions in S > m) $\leq (n(x-\varepsilon)-m)/(n-m)$, which equals $1+(x-1-\varepsilon)/f$. So we have that

P(I am f-wrong about S) = 1-P(fraction of propositions in S that are true > 1-f)

$<= 1- (1+(x-1-\varepsilon)/f)$

$= (1-x+\varepsilon)/f.$

Q.E.D.

Appendix B

The Brier score $B(P, D)$ of a probability function P measures the inaccuracy of P with respect to a finite set $D=\{X_1, X_2, ..., X_n\}$ of propositions. It is defined by

$$B(P,D) =df\ 1/n\sum_i (P(X_i)-T(X_i))^2,$$

where $T(X_i)$ is the truth value of X_i: 0 for falsity, 1 for truth.

Let P be the probability function of an agent who has good access to her own beliefs, in the same sense as described in Appendix A: the agent is certain that she has probability function Q, where for any X_i in D, $|Q(X_i)-P(X_i)| \leq \varepsilon$.

For a given value s, we seek an upper bound on how likely the agent can think it is that her Brier score is at least s. When $Q(X_i)=1/2$ for all i, the agent will be sure that her Brier score is exactly ¼. So we can only hope to get a substantive bound by assuming that $s > ¼$. Without loss of generality, assume that $Q(X_i) \geq ½$ for all i. (No generality is lost because we can always replace X_i with not-X_i whenever $Q(X_i) < ½$.)

Claim: $P(B\geq s) \leq (1-\surd(s))/\surd(s) + 2\varepsilon\ (\surd(s) – ½)/s.$

Proof: Notation: we write B for $B(P, D)$, P_i for $P(X_i)$, and Q_i for $Q(X_i)$. We proceed by computing an upper bound on the agent's expectation for her Brier score, and then applying Markov's inequality.

The agent's expectation for her Brier score is given by

$E[B] = E[1/n\sum_i(P_i - T(X_i))^2]$

$= 1/n \ \sum_i E [T(X_i)^2 + Q_i^2 - 2T(X_i)Q_i]$

$= 1/n \ \sum_i P_i + Q_i^2 - 2P_iQ_i$,

where the third equality follows because $E[T(X_i)^2] = E[T(X_i)] = P_i$, and because $E[Q_i]=Q_i$.

Give the agent's good access to her beliefs, this expectation can never be greater than when $Q_i = P_i + \varepsilon$ for all X_i. So let us assume this. Then we may continue simplifying as follows

$E[B] = 1/n\sum_i P_i + (P_i+\varepsilon)^2 - 2P_i(P_i+\varepsilon)$

$= 1/n \ \sum_i P_i(1-P_i)) + \varepsilon^2$.

Now, in order for the Brier score to be at least s, it must be at least s in the "worst case"--the case in which all X_i are false. The agent is certain that in that case her Brier score is $1/n\sum Q_i^2$. So in order for the agent to have any credence that her Brier score is at least s, it must be that $1/n\sum Q_i^2 \geq s$. Subject to that constraint, the agent's expectation for her Brier score is maximized when $P_i = \sqrt{}(s) - \varepsilon$ for all i, as can be shown by using the method of Lagrange multipliers.[13]

When P takes the above values,

$E[B] = 1/n \ \sum_i P_i(1-P_i)) + \varepsilon^2$

$= 1/n \ \sum_i(\sqrt{}(s) - \varepsilon)(1-\sqrt{}(s) - \varepsilon)$

$= \sqrt{(s)}(1-\sqrt{(s)}) + 2\varepsilon\ (\sqrt{(s)}-1/2)$.

In other words, whenever $P(B \geq s) > 0$, $E[B]$ can never exceed the above value.

Assuming that $E[B]$ achieves the above value, set $K = s/E[B]$, so that $s=KE[B]$. Then $P(B \geq s) = P(B \geq KE[B])$, which is no greater than $1/K$, by Markov's inequality (Motwani 1995, 46). So we have that

$P(B \geq s) \leq 1/K$

$= E[B]/s$

$= (\sqrt{(s)}(1-\sqrt{(s)}) + 2\varepsilon\ (\sqrt{(s)}-1/2))/s$

$= (1-\sqrt{(s)})/\sqrt{(s)} + 2\varepsilon\ (\sqrt{(s)} - \frac{1}{2})/s$.

Q.E.D.

Bibliography

Brier, G. W. (1950), "Verification of forecasts expressed in terms of probability," Monthly Weather Review, 75, 1-3.

Gilbert Harman, Change in View (Cambridge, Mass: MIT Press, 1986).

Joyce, J. (1998) A Nonpragmatic Vindication of Probabilism, Philosophy of Science 65, pp. 597-603.

Motwani, Rajeev and Prabhakar Raghavan. Randomized Algorithms. Cambridge; Cambridge University Press, 1995.

Roy A. Sorensen, Blindspots (Oxford: Clarendon Press, 1988)

Skyrms, B. "Resiliency, Propensity, and Causal Necessity." Journal of Philosophy 74 (1977), 704-713.

B. Skyrms. Causal Necessity. Yale Univ. Press, New Haven, 1980.

Talbott, William. "The illusion of defeat". In Naturalism Defeated?: Essays on Plantinga's Evolutionary Argument Against Naturalism by James K. Beilby (Editor) Cornell University Press (April 1, 2002).

Williamson, Timothy. Knowledge and Its Limits. Oxford: Oxford University Press, 2000.

Williamson, Timothy. "Probabilistic anti-luminosity," in Q. Smith, ed., Epistemology: New Philosophical Essays, Oxford: Oxford University Press, forthcoming.

[1] Thanks to Martin Davies, Campbell Brown, Ben Blumson, Ralph Wedgwood, Daniel Stoljar, and Ted Sider.

[2] P(student is named Sarah | overheard someone call her "Kate")

$= P(S|K)$

$= P(K|S)P(S)/P(K)$

$= P(K|S)P(S)/ (P(K|S)P(S) + P(K|{\sim}S)P({\sim}S))$

$= .05*.99 / (.05*.99 + .95*.01)$

$\approx .84.$

[3] P(student is named Sarah | overheard someone call her "Kate") =

$= P(K|S)P(S)/ (P(K|S)P(S) + P(K|{\sim}S)P({\sim}S))$

$= .05*.90 / (.05*.90 + .95*.10)$

$\approx .32.$

[4] Cf. example 3 ("The Antidote") from Talbott (2002), 160-1.

[5] P(memory is reliable | seem to remember doctor saying "your memory is unreliable")

$= P(R|C)$

$= P(C|R)P(R)/ (P(C|R)P(R) + P(C|{\sim}R)P({\sim}R))$

$= .01*.9 / (.01*.9 + .2*.1)$

$\approx .31.$

[6] In an even worse case, AE believes not just that his confident navigational judgments tend to be wrong, but also that he only suspends judgment in cases in which the thing to do is turn left. Here AE's beliefs are unstable even if he suspends judgment.

[7] For a discussion and defense of this claim in the case of all-or-nothing belief in a single proposition, see Sorensen (1988), 392-396.

[8] We note that this claim is disputed. See, for example, Harman (1986).

[9] Compare the discussion of "Most of my beliefs are false" in Sorensen (1988), 48-49.

[10] Why can't an agent reasonably come to believe that she is an anti-expert simply by conditionalizing on the proposition that is an anti-expert? Because conditionalizing on that proposition would make the agent very inaccurate about her own beliefs, and would thereby make her unreasonable.

[11] The Brier score of a probability function P on a list of n propositions $\{X_i\}$ is defined to be: $1/n\sum_i(P(X_i)-T(X_i))^2$, where $T(X_i)$ is the truth value of X_i: 0 for falsity, 1 for truth.

[12] Compare Sorensen (1988), 391-2.

[13] We wish to find $P_1,...,P_n$ which maximize $f(P_1,..,P_n)$ subject to the constraint that $g(P_1,...P_n)>=0$, where $f(P_1,...,P_n) = E[B] = (1/n\sum_i P_i(1-P_i)) + \varepsilon^2)$, and $g(P_1,...,P_n) = (1/n\sum_i Q_i^2) - s = (1/n\sum_i(P_i+\varepsilon)^2) - s$. We may instead use the constraint $g(P) = 0$, since increasing $g(P_1,...,P_n)$ can only done by increasing some P_i, which reduces $f(P)$. Extrema of $f(P)$ occur when $\text{grad}(f) = \lambda\ \text{grad}(g)$, i.e., when for all i, $\partial f/\partial P_i=\lambda\partial g/\partial P_i$. Substituting, we have $1/n(1-2P_i) = \lambda/n\ (2P_i+2\varepsilon)$, from which it follows that $P_i = P_j$ for all i, j. So we have that $0 = g(P) = (1/n\sum_i(P_i+\varepsilon)^2) - s = (P_i+\varepsilon)^2$. It follows that $P_i = \sqrt{(s)} - \varepsilon$ for all i. A check shows that this is indeed the solution that maximizes E[B] subject to our constraint.