Rational Self-Doubt: The Re-calibrating Bayesian

If one is highly confident that #3 in the line-up is the murderer from having seen the crime, and then learns of the substantial experimental psychology evidence that human beings are unreliable at eyewitness testimony, is one thereby obligated to reduce one's confidence? How far, and why? I generalize 1^{st}-order Bayesian rationality constraints away from idealizations in a principled way, to give a rule for revising 1^{st} order beliefs on the basis of 2^{nd}-order evidence about one's reliability. It is a conditionalization rule that re-calibrates the subject and sidesteps standard objections to calibration. It shows why taking doubt about one's own judgment seriously does not end up in incoherence or runaway skepticism, and what the added value of this kind of evidence is. I'll discuss some applications to disagreement and testimony, and some preliminary results from its implementation in an AI program.

SLIDE 1: This is the central piece in a larger project that began with questions about how to cope rationally with the fallibility of science, and with the growing body of results of empirical psychology – for example, about eyewitness testimony – that tell us human beings are not nearly as reliable as we tend to assume of ourselves.

These kinds of evidence are general, and the question is how they should that affect an individual's confidence on a specific case of the same kind of question. What if you're very confident that the murderer is #3 because *you saw him*, and then you hear about the psychology literature.

SLIDE 2: So what are we supposed to do when we acknowledge fallibilities or even unreliability to ourselves?

- One possibility is just to observe a moment of silence. That seems pious but also a little thin.
- Should we instead revise our confidence? – Well, if so then by *how much*, and why exactly?
- In responding to our fallibility, on the one hand, we don't want to be immovable for the sake of it – although this is sometimes advantageous in politics.
- But on the other we don't want to be what we call *indecisive*, or *wishy-washy*, or *hand-wringing*. [*Chronic* self-doubt is a symptom of many conditions considered mental illnesses.]

  And it does seem easy to fall into that spiral if we start doubting our judgment at all. Since isn't the same self-doubt always available *again*? How do we stop and settle on what to believe?

Are there *rules* for ending it? You *can* end it just by saying I need to act now – maybe that's what the healthy person does. But that looks arbitrary *epistemically*. Is there a way to be both healthy *and* rational?

- There's also an intuitive problem here about the unity of the self. How can you be consistent and a single self in the moment when you recognize the conflict – and in some sense disapprove of your own belief?
- I think the epistemic question here about what to do with general information *about our beliefs* when it seems to threaten the justifiedness of a particular belief is actually a hard problem, and that may be one reason average "healthy" people don't take as much account of our fallibility as maybe we should.

SLIDE 3: Examples of evidence of our fallibility are everywhere, and variable in type. READ EXAMPLES.

- The evidence may be specific to the human being as in the first and 3rd – evidence that you *in particular* took a mind-altering drug, evidence that your *personal* visual system isn't working – or it may be about human beings in general of which you are one, as in the second and fourth.
- Different types of evidence; In the third you have a single indicator, the change in the visual field, In the second you have the results of a large empirical study, in the first testimony. (The fourth is actually not an established result.)
- You also have different possible directions of revision. In the first three cases it looks like you should dial your confidence down, and in the fourth up.

There are the variations but all of these cases seem to have something in common. At the very least they are all cases where it seems there's pressure to disapprove of and to revise one of your own particular beliefs. So this topic is the *complement* of what Bob Stalnaker was concerned with yesterday. He talked about cases where we ought to endorse our own beliefs. You could say what I'm talking about is the dark side. But if what I propose works then we *can* doubt our own judgment without falling to pieces or becoming inconsistent. There's a way of responding *proportionally*, so it's not a night in which all cows are black.

SLIDE 4: In those cases there was straightforward pressure to revise, but there are cases that look very similar where it seems we shouldn't …

- The evolutionist admits to the Creationist that our theories *might be wrong*, and then all hell breaks loose. The Creationist concludes his view is just as good. Every view is just a hypothesis! Now there's definitely something wrong with this argument. But *what*?

- The pessimistic induction over the history of failed scientific theories is somewhat similar in appealing to general evidence of fallibility and expecting a drop in your confidence in your particular theories. But even though we hope the pessimist is wrong it is prima facie worrying. He might seem like a case that belongs on the previous page. *What's the difference between him and the creationist*?

- You are contemplating marriage and read that the divorce rate is 60%. Maybe people should think more about the divorce rate than we do, but it's hard to believe that all 120 million Americans currently married were flat irrational to take their vows in full knowledge of this statistic. Problem is why aren't they irrational? [The assumptions behind 3 being a problem of course are that promising has epistemic requirements and the promise made in marriage vows is literal.]

- So, we need guidance about why there should be different responses to different cases, and in some cases no response at all.

And what I've been looking for and formulating is a general rule justified via more general and already accepted *principles* that will answer the question what is the rational way to revise (or not) in these widely varying cases, and will give an understanding of *why* those are the answers.

It will soon become obvious that this work is incomplete and there are possibilities and issues to address that I'm not aware of, so I'm hoping for patience and help.

In that spirit, I'm going to start very basically, where the first thing to notice is that despite the variation in types of cases, the evidence that's coming in is *about* your beliefs and their general relationship to the truth.

SLIDE 5: It has to be distinguished carefully from ordinary evidence for or against the *content* that you believe.

Say that C is the proposition that the convection currents of the Sun have a certain effect on the rate of solar neutrino flow to the earth. e is data about this modeled appropriately. p is the observation that, say, 80% of past claims about unobservables that were supported by apparently good evidence are now known to be false.

e is evidence about the Sun. It is *1st-order evidence*.

p is *not* evidence about the Sun, but it is evidence apparently relevant to the reliability of your belief about the Sun because those convection currents are not observable in the intended sense.

p provides no direct reason to doubt C, but it does give reason to doubt the fitness of your judgment of C *via* what it says about the means by which you formed a degree of belief about C. Since it is evidence about beliefs, it is *second-order* evidence.

SLIDE 6: You might think, second-order is *complicated* especially if we go to probability. Can't we just express all of what we want to say at the first order? We'll see later that we can't if we want to say what we need to say and be coherent. But intuitively, compare the following two cases to see why we can't explain or even express at the first order what we need to say.

You are highly confident of q – that there's no tiger around – and then see an orange, furry rustling in the trees. *vs*. You are highly confident of q, and then your visual field goes blank (uniform).

- In both cases you should withdraw confidence that there is no tiger around.
- But in the second although you got new information, you have *no new evidence about tigers*. As initially, you have no pixels indicating tigers, so

the most it can say about tigers is *no tiger* (!). But obviously that's not what you should think.

- There's nothing at the first order (in the pixels) to explain why you should worry that there's a tiger around, so this can't be modeled as learning from first-order evidence.
- The key is that in order to state the relevance of the uniform pixels to whether you ought to believe there is no tiger, you have to *refer* to a belief or beliefs, as in "Beliefs formed via my indicators are not related to the truth anymore." Employing the word "belief" makes your statements second order.

SLIDE 7: We have to express 2nd-order evidence more precisely. Whatever is evidence or hypothesis for you ends up as a belief (if it's going to be able to prompt changes in your other beliefs). So second-order evidence is a belief about one or more beliefs. And this is how those are built and expressed, in words, to just go through the syntax …

Say you have a proposition that doesn't contain a belief predicate – the murderer used a knife. If you believe that, you have a first-order belief. If you believe that you believe, that's second-order and you express that via a sentence that has a belief predicate in it. You say "I believe that the murderer used a knife."

– I ascribe a 1st-order belief to *another* person with one belief predicate by saying "He believes q" and I ascribe a second-order belief to him by using two belief predicates nested, saying "He believes that he believes q." In that last case I'm expressing a third-order belief.

"I believe q." is literal, not a hedge, not an intensifier.

Sometimes in ordinary life "q" and "I believe q" are used interchangeably or with "I believe" as a qualifier indicating less confidence (hedge) – as in "I *believe* that the murderer used a knife" – or as an intensifier (as in church). Here I'm reading it literally, a belief about a belief. It will look easier when we use nested functions.

SLIDE 8: To give us a quantitative handle on the question I'm going to talk about not beliefs but degrees of belief, or degrees of confidence, and

- I'll express all the talk about degrees of belief in terms of probability in the usual way. But we'll also include sentences that are themselves statements of probability, i.e., statements that someone has a certain degree of belief in some proposition.
- Now we're going to have a subject have a degree of belief about a proposition that states some subject's degree of belief in something. This is expressed by composition of probability functions. You would read the first sentence as Tonya's degree of belief that Sergio's degree of belief in q is x, is x'.
- Using the "P" means the subject's degrees of belief conform to the probability axioms
- A subject will be evaluating her own beliefs in the cases that concern me, so the main issue is second-order probabilities with a function *applied to its own statements*. Sergio is x'' confident that his degree of belief in q is x.

SLIDE 9: The simplest objections I face are to using second-order probability or second-order belief at all, but fortunately Brian Skyrms addressed a lot of them a long time long ago.

First was an idea that higher order beliefs didn't really exist. Assertions of probability are expressions, not true or false. But this is easy to answer: an expression is a state and you can ascribe states to a person. We can give the usual Ramsey line about beliefs at the second-order. We could measure your beliefs about what your beliefs are by bets. We can ask you to bet about how you would bet.

Another objection I sometimes still hear is that second-order probabilities are *trivial* because they'll all be 0 or 1, and hence strictly irrelevant to everything. You would think that if you thought we had perfect knowledge of our own beliefs.

But someone can be very sure they don't have racist beliefs and then fail an association test. You might think that means they're wrong about their beliefs. I think we're not infallible about our beliefs (although I actually won't need to appeal to that assumption in my construction).

But one could reply to this point about fallibility by saying that Bayesianism is an idealization. We are also not actually *consistent* but we tend to like rationality constraints that require it. Maybe the *ideally* rational agent has perfect knowledge of her own beliefs.

I think that can't be right if you want to stay within a Bayesian-inspired system since to what extent you believe q is a contingent matter of fact – in respect of contingency it is like the number of ice cream cones in San Francisco – and Bayesianism proudly lacks requirements on substantive knowledge. So I think it's not only permissible but a positive thing to get away from that assumption if we can, and when it's useful to.

SLIDE 10: To give a general answer to the question how to assimilate this higher-order evidence we have to say something about how first- and second-order beliefs relate. I'm going to generalize a standard bridge principle between first and second order beliefs, which will give us both synchronic and diachronic constraints on dealing with second-order evidence. Then I'll defend them and so on.

SLIDE 11: Here is the principle that apparently gets in the way of understanding what we should do with this type of evidence that pressures us. Your degree of belief in q given that your degree of belief in q is x should be x. The degree of belief you think you have in q is what your degree of belief should be.

Don't disapprove of your own degrees of belief.

- Looks right – many have thought it trivially so, undeniable.
- But it also looks like the earlier examples might make it false.

Seemed it was possible there to both see that you *have* a belief and see that it's not what you should have. But it seems that would be expressed by having the first x not equal the second.

- MP is a very tight bridge principle. In particular it is impossible to disapprove of your own beliefs.

- Note that this is a bridge principle so it is independent of the axioms (which are within-order constraints). This demands a relation between the first and second order.
- So whatever we say here we need arguments that go beyond the axioms.

There are conflicting intuitions about this principle – that it is undeniable and that it has counterexamples – and I think the reason is that we're responding to more than one version of it.

SLIDE 12: We need to distinguish between what I'm going to call restricted and unrestricted self-respect.

Restricted says your degree of belief should be what you think it is provided there is nothing to tell you it should be otherwise.

RSR is surely unobjectionable, no reason for disapproval is in the condition – The mere fact that you have a degree of belief is not by itself a reason for it to be different.

but in our examples *we were not contemplating changing our confidence in q merely on the basis of having noticed that that was our confidence*. So, those are *not counterexamples* to THIS principle.

SLIDE 13: Which means that this is a principle we want to preserve in finding a more general bridge principle.

SLIDE 14: Our examples are counterexamples only to *this, stronger* principle. [Skyrms knew this was false. He described counterexamples, but they were pathological, and it was very useful and legitimate to ignore them because interesting things can be proved with this.]

Your degree of belief should be what you think it is no matter what else you believe. I'm going to tell you now, by one argument,

- WHY we don't want this principle
- WHAT we want instead
- HOW that thing preserves RSR and generalizes it

SLIDE 15: I'm going to develop this by just expressing the kinds of claims we saw in the examples earlier *literally* and *explicitly*.

- Reliability as defined here depends on the proposition, degree of belief, and person, because as a matter of fact our reliability varies over all of these dimensions.
- This expression says your reliability y is the objective probability of q given that you have confidence x in q. That may seem weird because a low confidence from you could count as highly reliable, but
- this expression is what you get when you ask How far is the fact that you believe q to degree x an indicator that q is true? More explicitly, it says to what degree the subject's having confidence x in q *confirms* q, and combines this with a prior on q or P(q)=x. It equals how far x – whatever it is – takes q to being true. How far it is evidence that q is true. It's measuring both what Carnap called "incremental" and what he called "absolute" confirmation. And the fact that you have a 20% degree of belief could be an indicator that q is likely true, if you're an unusual kind of person who is largely anti-correlated with the truth. If we knew that about you, then we could figure out that q is probable via your low confidence in it.

SLIDE 16: There are several ways of seeing what information is in this reliability term. We can spell the expression out via Bayes Thm, or via my leverage equation. (*TT*, Ch. 5)

Either way, to write it down imagine that q is the hypothesis and the statement that you have a particular degree of belief is the evidence.

In the first way (via Bayes Theorem) you get to the reliability expression by asking how far the subject's degree of belief indicates q – via the ratio measure of confirmation -- and combine that with a prior on q.

In the second case (via the Leverage Equation – *Tracking Truth*, ch. 5), you express the same reliability expression via the ratio measure and the likelihood ratio measure (and a prior on the evidence is in there).

So, you have both priors (base rates) and quality of evidential support (type I and type II error) being evaluated in this term.

SLIDE 17: So this reliability term is a result of measuring something epistemically interesting. Also the likelihood ratio has a correspondence with the tracking conditions on knowledge, and the direction of fit in the term itself corresponds to safety – that q be unlikely to be false when you have some degree of belief in it.

SLIDE 18: To get a further sense of what this reliability term means, it is a function specific to a person, that gives a value for every confidence: when you believe q, a certain type of proposition, to degree x, then q is true y% of the time. Psychologists measure this in human subjects, and the functions graphed are called calibration curves, and it's more or less the same concept as used in weather forecasting. The green curve is perfect calibration where your confidence always matches your accuracy and the red one is a common one you find with human beings. We tend to move from underconfidence to overconfidence as questions get more difficult, with the curves crossing at 75%. (Interestingly, a C on an A through F scale.)

SLIDE 19: Now to express the quandary that all of those examples give us, we just write it down: We want to know what a subject's degree of belief in q should be given that the subject has degree of belief x in q and the objective probability of q given that the subject has degree of belief x in q is y.

SLIDE 20: Visually easier:

SLIDE 21: Now, Unrestricted Self-Respect tells us to ignore the information about our reliability which now appears in the "r" place. Stick to your guns. Nothing is a reason to have a different degree of belief than the one you take yourself to have.

SLIDE 22:

But if you look at the syntax of this you see that the LHS is a substitution instance of the LHS of the Conditional Principle, if we take credences to be probabilities and restrict PR to chance.

SLIDE 23: That principle, related to the Principal Principle, says that your credence in q given B and that the chance of q given B is y should be y.

The Chance term is our reliability term if we substitute for B the statement that the degree of belief in q is x.

SLIDE 24: This means the Conditional Principle gives us the answer y, not x. As far as I can see this conditional principle hasn't met with any objections. (See Vranas 2004, PPR, "Have your cake and eat it too: The old Principal Principle reconciled with the new") Discussion of admissibility conditions would carry over as they already are, with of course the possibility of new wrinkles, and even problems for me.

SLIDE 25: The answer USR gave us was x, not y. And the question becomes which principle we should go with.

[If instead of PR you had P, and if x did not equal y, then you would be first-order incoherent. So that term needs to be objective, or at least a different probability function from P.]

Well, USR has apparently very clear counterexamples. The CP is very intuitive and no one has raised objections. So until I see a problem (and maybe even after that) I'm going with the conditional principle. Of course what I'm about to do may make you start worrying about the Conditional Principle. (Notice that I'm using PR instead of chance, hence the star. That's just because I would like the principle to be as general as possible over objective interpretations of probability. I don't currently know how far that is.

So this discussion has made a connection between our *intuitive* questions and bridge principles between subjective and objective probabilities.

SLIDE 26: Another argument for liking this option over unrestricted self-respect is a symmetry argument: our respect for the judgment of others is not unconditional, and CP agrees. (which you can see by taking the term B as $P_S(q) = x$.) Given that we know we're not infallible, why should we behave differently toward ourselves?

This first equation gives an intuitively very natural way of regarding other people's beliefs:

In forming her own confidence about q, surely Tonya takes (or should take) Sergio's testimony that q exactly as seriously as she regards him as reliable.

Her degree of belief in q given that sergio's degree of belief is x and the objective probability of q given his degree of belief is x is y, should be y.

[Note: this is crude. We'll see later that this says Tonya should take Sergio's confidence of x to give her a confidence of y if and only if she is certain that Sergio's confidence x is a perfect indicator that q is y-probable. I endorse this (modulo that in this case she has to be uncertain or (even slightly) wrong about her degree of belief), but we'll Jeffrey-ize the whole thing because we never have a right to be certain of someone's calibration curve (except, hypothetically, in the infinite long run). This is because the calibration curve is what's called an empirical fact and I'm what's called a "radical" probabilist, which isn't radical at all but just means non-Cartesian.]

SLIDE 27: So the rule that I've expressed for *self*-doubt is an instance of a *general* principle about how to regard people's beliefs. It is the instance S = T. Our way of dealing with our own reliability is a special case where Tonya applies this reasoning to her own probability function.

This accords with the idea that self-doubt requires "getting distance on yourself" but still treating yourself as a person. You should treat yourself just as you would anyone else. [The challenges are going to be whether we can do this consistently and what the added advantage of doing it is.]

[So, clearly since we now know how to represent the subject's ascription to herself and others of beliefs and reliabilities, we can represent a lot testimonial situations, including those where we compare our beliefs and reliabilities' to others'. If we represent peer disagreement as my conditioning simultaneously on my having certain beliefs and someone else having different ones, and the two of us having the same reliability, then the condition ends up being a contradiction and undefined. (See 2009, "Second-Guessing: A Self-Help Manual.") If these ascriptions to myself and the other come sequentially then it's not undefined, but the answer depends on further factors, such as how good my evidence is about myself and the other, and is governed completely by Jeffrey conditionalization.

SLIDE 28: So the derivations lead me to two principles, one telling you how your beliefs should relate to each other at a given time, and the other how they should relate over time, that is, how you *learn*.

- READ Cal: supposing that thing (in the condition) is what you believe about yourself, then y is what your degree of belief in q *should* be. [Note that this state is one in which you are acknowledging the possibility that your degree of belief isn't what it should be.

- In Re-Cal I've just made this diachronic, by assuming that conditionalization is a good way to learn. If x doesn't equal y then what you believe about yourself (there in the condition) implies that you are out of line with the CP. By putting the initial and final subscripts there what we have is a rule of conditionalization for *how to get back in line with the conditional principle* when you have evidence you've fallen out.

When your confidence matches your believed reliability we're going to call you calibrated, understanding that …

SLIDE 29: this is a s*ubjective* version of calibration.

This is a strict formulation, but where $x \neq y$ these will mostly be Jeffrey conditionalizations, which I'll get to.

SLIDE 30: The Re-Cal principle explains the earlier cases very nicely I think.

The reliability term in that equation is the main thing that tells you how *much* you should revise your confidence and in which direction, and the various kinds of higher-order evidence should all affect your estimate of that reliability term.

- Eyewitness case: The psych evidence is relevant to your *reliability*. You have to consider all the evidence to figure out how these psych conclusions bear on your case. But if you know nothing to mitigate your averageness in this, then the psych evidence is what you have to go on in judging your reliability. E.g., switch from confidence .99 to say .70 [In fact, there will be a lot of mitigating factors in every case, as I'll get to, but this is the idea.]
- Tiger: You don't learn about the tiger but in that single cue in which your visual field goes uniform you do learn about your reliability. You learn that the thing forming your belief is not an indicator either way. PR(q/P(q) = high) = .50 because that's your new estimated reliability.
- The difference between the creationist and the pessimist is that: *quantity matters*: admission that you might be wrong is only admission that your reliability is <1. That only means that your confidence in your particular theory needs to be less than 1, i.e., not certain. Scientists are already there so no change is required.
- Pessimist: typically assumes that a very high percentage of the scientific theories of the past have been wrong about unobservables. Now that's a serious degree of unreliability of people who were "like us." That says that because you are doing what they did and they were confident like you, the probability you are right given that you are confident is less, possibly much less, than 50-50.

I don't think that argument works, and you can read about that, but this does explain nicely why it is prima facie threatening, which I think we have to grant.

In the case of the woman we can explain why the confidence should go up. It is perfectly possible for the y to be greater than the x, for your reliability to be greater than your actual confidence. Possible that you are a person whose hesitation is greater than it should be given your reliability. [The statement itself, though often what people think on anecdotal evidence, is not an established fact. The relation between gender and calibration is open, with mixed results in studies.]

SLIDE 31: In the marriage case, there are probably people who have not taken into account any information more particular to themselves than the general divorce rate. In that case they are irrational if they make the literal promise “til death do us part” because they have good reason to believe they can’t keep it. But there are people who think about the divorce rate and bring up particular facts about themselves – we have an effective way of dealing with conflicts, etc. “That won’t happen to us!” “We’re different!” isn’t information but there can *be* information. There are also people for whom the evidence specific to themselves makes it obvious they are *worse* than average and so definitely irrational to get married if the promise is literal, e.g. Larry King.

SLIDE 32: Read slide.

SLIDE 33: Intuitively, it is puzzling how a person can doubt her own judgment, and remain consistent and one subject. I take is as a virtue of my view that it can represent and prima facie explain these things. But keeping it one subject – which I represent by making a probability function apply to its own statements -- also leads to more worries about consistency than you would normally have with second-order probabilities that use different functions. So we have to address this issue explicitly.

SLIDE 34: There are at least three kinds of concerns about incoherence.

1) Against Cal and Re-Cal in particular.

2) Against applying a probability function to its own statements.

3) Against the need for re-calibration at all due to the relation of calibration to coherence.

SLIDE 35: what if …? Read

- 1, 2 amount to assuming perfectly accurate and complete self-knowledge of what your degree of belief in q is.
- 1 and 3 are the condition in Cal so from those we discharge to get P(q) = y. But by 2, P(q) equals x, and by 4, $x \neq y$. $\perp$
- Is this really a problem?

SLIDE 36:

- If we subtract even epsilon from those certainties in 1 and 3, then this argument doesn't go through.
- So the extreme probabilities are doing the work.
- Role here is they artificially induce independence relations, in particular that the reliability term is not relevant. [Those independence relations can't be assured another way without making the subject's degrees of belief completely independent of what she thinks the objective probabilities are (see 2009 Second-Guessing), so there's no way to rescue this argument.]
- What about the case where we *do* have certainties? If you are certain and accurate about your degree of belief and certain of your reliability, that forces x=y, that is, it means you are subjectively calibrated. There's nothing wrong with that, and if you're there then of course Re-Cal won't force you to change anything.
- But I can grant perfect self-transparency about what your beliefs are, and still have you in need of re-calibration because it is not going to happen that you have enough evidence to be certain of what your reliability term is. This is a substantive matter and our evidence is always incomplete. So premise 3 will always be false.
- All of this does mean that most Re-Cal conditionalizations – all those except ones failing in accuracy – will be Jeffrey (not strict).
- Note, if x does not equal y then you are in the condition regarding yourself as in violation of the Principal Principle, but that's not a violation of the axioms, not a case of incoherence. It's not even regarding yourself as violating the axioms, or actually violating the Principal Principle.
- Can that independence be insured in other ways? If there is that independence then it follows that her subjective probabilities are independent of what she thinks the objective probabilities are. That kind of subject would have way more problems than fallibility. I'm okay with the consequence that Re-Cal can't help her.

So Cal allows lots of options for remaining consistent: imperfect self-knowledge, imperfect knowledge of your reliability, or you are subjectively calibrated already – and these situations are sufficient for the rule to do its job for me.

SLIDE 37: A second consistency worry is addressed to the fact that we're applying probability functions to probability statements.

If you allow the application of a probability function to its own statements, then, Skyrms pointed out, a power set paradox can be generated by associating with each subset of the domain a distinct proposition, giving you a mapping from the subsets of the domain into that domain.

These are Boolean combinations of things that are propositions, so they are also propositions.

- 1-1 mapping of the power set of the set of propositions into the set of propositions.
- Nesting of one probability function on itself allowed this.
- If we had *different* functions, different domains, never their own statements or the statements from bigger domains in their own domains, a hierarchy, so no one domain could take its *own* power set.

SLIDE 38: Skyrms replied to this by using a typed theory. So, you never have a probability function applying to its own statements, but only one probability function applied to the statements of another. That way the domain of any function is always outstripped by its power set.

On this formulation Re-Cal would have the function in the condition be different from the higher order one, prime vs. double prime

SLIDE 39: But a typed theory won't be adequate for my purpose.

We have two functions, but they both have to be me. Prime me is in the position in the equation to be the one that double prime me is observing and basing her new confidence in q on. She will often come up with a different confidence in q than prime me had (due to having run prime me's confidence through an estimate of prime me's calibration curve). Double prime me thus has the ability to *disapprove* of prime me's opinion. However there is no provision for her to *correct* prime me's confidence. Also, I am now a divided agent in a very straightforward sense. If you ask me to bet on q, which probability function will answer, prime me or double prime me? This is why if you do use two functions you need that tight bridge principle that has been so popular, and that forces the levels to agree.

- But since I'm applying one probability function to its own statements, I still owe an answer to the Power Set Problem.

[Hannes Leitgeb: You don't need to make such a fuss about the consistency problem. If you use modal logic people put K on K without worrying. And you should look at Gaifman, probability functions on probability functions.

Response: It's true using modal logic would be convenient for presentation because applying K to K is represented with an easy picture, and it doesn't give people the willies. But 1) I want to keep potential inconsistency in my face because it corresponds to a big part of the intuitive problem, 2) I've checked and modal logic gives inconsistencies at the same endpoints that we have here, and 3) my account has to answer the quantitative question how much to adjust, so I have to put a probability function on the modal logic anyway. On Gaifman, as far as I know he and the expert function literature don't apply probability functions to *themselves*.]

SLIDE 40: But as we know a typed theory isn't the only way to address set-theoretic paradoxes. My solution: the class of propositions is not a set. And one thing I do know is that probability is definable on proper classes. If anyone knows what the costs might be for making this move I'm all ears. But yesterday I learned that people work on type-free probability so I think I can learn something there too.

SLIDE 41: There is another concern about coherence. There's a variety of great work out there that might give you the impression that adherence to the axioms – first-order coherence – already gives you calibration. We need to go into some details to see why that's not true enough to ban recalibration.

SLIDE 42: What van Fraassen showed for example is that coherence implies not calibration but that it is not *a priori impossible* for you to be calibrated, leading to a defense of the axioms analogous to a Dutch Book. (Of course, here I'm not arguing for the rationality of adherence to the axioms but taking that for granted.)

SLIDE 43: In other work by A.P.Dawid and Teddy Seidenfeld it is shown that the Bayesian must regard himself as someone who will be calibrated in the long run. It doesn't yield even subjective calibration at a finite time, but only the option of consistently ignoring a finite set of evidence that it would be natural to regard as evidence of miscalibration. Coherence *alone* doesn't guarantee short-run calibration and doesn't forbid you from regarding yourself as uncalibrated at a finite time and re-calibrating.

SLIDE 44: Thus coherence leaves you options in the short run.

In fact, the reason the subject must believe that he will be calibrated in the long run is that he can't ever learn from a subjectively zero probability event (his long run miscalibration), which kind of makes that "convergence" to subjective calibration look worryingly trivial.

SLIDE 45: This is also stated explicitly in the paper in which Dawid discussed the result. He is *doubtful* that there is a coherent way to do short-run re-calibration but clearly says that the theorem has not shown it to be impossible.

SLIDE 46: But he thinks it's obvious that one should re-calibrate short-run so takes this theorem and inability to formulate a Bayesian rule for finite re-calibration as a big mark against bayesianism.

A way to see why Dawid thought it obvious that you should re-calibrate is to think back to our original examples. The supposed "Bayesian" recommendation not to recalibrate is the original response I called pious but thin: stick to your high confidence that the murderer is #3 and simply ignore the mountain of evidence that human beings are overconfident in eyewitness testimony, *even if* you have no other evidence that personally excuses you from that general trend. *It's okay*: as long as you're coherent you'll be calibrated in the *long run*.

Dawid couldn't see any detailed rule or principled argument for short-run re-calibration – and showed it is a hard problem. You could see what I've argued here as proposing a bayesian answer to just those questions (although highly abstract, not addressing some of the statistical problems Dawid brought up). Notice though, I'm well aware I haven't given a PROOF that Re-Cal preserves consistency, but only addressed particular arguments that it doesn't. Though the proof where extreme probabilities introduce independence does generalize against any argument where the assumptions introduce independence. See Roush 2009.

Dawid thinks it's obvious that one should be concerned about short-run calibration, but he doesn't give a principled reason why. I'm giving the Conditional Principle. It's not that I think I've fully explained why we should aim for calibration, but I have reduced it to a question about this principle.

SLIDE 47:

So if all of that works, we're 4 objections down, three to go. Next is the worry of regresses.

SLIDE 48: In following this rule you start with a first order confidence in q and end with a new first order confidence in q, which you can notice as easily as you did the one you started with. Doesn't that mean the rule is applicable again? How is a regress avoided?

Note first that this question has a nice correspondence to the potential for a person to not know how to stop second-guessing their judgments and people who do *that* don't know what confidence in q to settle on. I think it's an advantage of the account that it gives an exact model of what chronic self-doubters do wrong.

[Also corresponds to what people think happens when we think of justified belief as requiring that you have a second-order argument available to the effect that your first-order belief was formed in a good way. Both of these are mistakes.]

So why doesn't this rule yield a regress? First, maybe obviously, notice that the rule does not demand that you complete some infinite hierarchy to get back down to the first order, a worry that people have voiced about higher order probability. That's a misunderstanding. You go to the second order and back to the first, and that's it.

But when you get back to the first order, isn't the rule applicable again? So isn't there a regress? How do you stop?

SLIDE 49: This is a conditionalization like any other. You're not allowed to do it – indeed you have nothing to do it with – unless you *have new evidence*. If you don't have new, then you just stop and go with that degree of belief. We don't think of 1-st conditionalization as yielding a regress and we have no *more* reason to think that here.

SLIDE 50: And notice that the evidence that you need to go forward in doubt after you go from x to y is about what your reliability is when <inhale> [your degree of belief is y *and you got that by re-calibration from* x]. The evidence you used to re-calibrate your first degree of belief x is (not only spent but also) not directly relevant to this new question. To the extent that it is relevant it's already been taken into account in getting you to confidence y.

(As evidence already conditionalized on it is probabilistically *irrelevant*. I say "not directly relevant" only to acknowledge the intuitive sense of relevance, that that evidence about confidence x was used to get you to new confidence y.)

There's necessarily going to be less of the type of evidence around that you would need for a 2^{nd} re-calibration, namely as to how accurate you were when you "have y due to a re-calibration from x)" than there was for how accurate you were when you had x, because it has more conditions on it. In actual cases that evidence peters out very quickly. And I think this is the mistake chronic self-doubters make, of not realizing that they have no evidence for the further spinnings from the first-recalibration to more re-calibrations.

- So this is no more a pathological regress with this 2nd-order conditionalization than the first-order conditionalization rule proposes. It is simply a rule for how to respond to evidence as it comes in. Stopping when the evidence stops is not epistemically arbitrary.
- As always, in conditionalization, the question is what is the right degree of belief *given the evidence you have*.

  People often seem to have the worry that when you do this second-order conditionalization you're throwing out your first-order evidence that got you to the degree of belief you're now correcting. Not at all. The correction you're now doing *depends* on the value of the x that your 1st-order evidence got you to. That first-order evidence tells you which bin to look in to get the correction. It's a different correction depending on which confidence you're correcting (which bin you look in). Whether the correction is a full replacement depends on the quality of the 1st-order as opposed to second-order evidence, which is taken account of in the fact that you have to do this conditionalization Jeffrey style. (Talked about below.)

SLIDE 51: Next in line in objections, Teddy has argued that short-run recalibration is distorting. The problem is: You can be calibrated in your confidence about rain by knowing that 20% of the days in the year it rains in your locale and announcing 20% chance of rain every day.

You are perfectly calibrated but have no discrimination, you say nothing more specific about particular days. You are not as informative as people we can imagine who are less calibrated, say consistently overestimate by 5%. You could hedge your bets this way and ignore more specific evidence you might have in order to get the highest possible calibration score.

→ Calibration is an improper scoring rule.

SLIDE 52: The first thing to say is that Re-Cal is not a scoring rule, but a principle of conditionalization.

What I'm imagining is not a game where you announce a confidence to someone and get scored and try to maximize calibration however you can. In the question whether you carry out a principle of conditionalization there is no one to lie to.

In a conditionalization there's also no choice at all about HOW to re-calibrate. Conditionalization rules give a unique answer to what your new degree of belief should be, based on the evidence you have and your priors. In using a conditionalization rule you maintain attention to the epistemic virtues that Bayesianism enforces.

Hedging your bets isn't possible because the conditionalization rule is over your whole probability function and so completely enforces the principle of total evidence. The only way you could be allowed to achieve perfect calibration by moving to a 20% confidence in response to information that 20% of the days in the year are rain days is if you have *NO OTHER INFORMATION*. That's not hedging. It's ignorance, not irrationality.

Anyway, I don't claim that calibration makes up for lack of information. (That would be silly.) So a better comparison is between two people who have the same amount of information and one is calibrated and one is not. If a person is calibrated his confidences match the true probabilities, so he'll have more success in betting. (Alternative presentation) Two cases: Either he has more evidence about rain tomorrow than the yearly statistic or he doesn't.

- If he does then he is hedging in saying 20% every day, but he's also violating the principle of total evidence. (Conditionalization enforces not violating it.)
- If he doesn't have more info about rain than the yearly statistic, then there's surely no shame or hedging in setting his confidence at 20% every day.

SLIDE 53: Here's another fair question that comes from conversation: Say you have *one* data point about your reliability. Are you seriously saying it is rational to update on that basis?

SLIDE 54: Well, this question arises at the first-order too and there are a variety of ways of dealing with it. In implementation for example you can set your learning rate to vary directly with size of the data sets, so that you reduce the effect of small data sets on your degree of belief. Small data sets can always be distorting, but since everyone thinks you can deal with at the first order it's reasonable to expect you can do similar things at the second order.

SLIDE 55: The fact that this can be controlled is shown in the Jeffrey form of Re-Cal where you don't learn your degree of belief and reliability with certainty. You can see in this more explicit formulation that there are a lot of questions to be answered in deciding how far to change your degree of confidence.

SLIDE 56: The bold black term is just that reliability term given by the Conditional Principle (CP), and its value is y. The red term, what your confidence in q should be if these things about yourself are not true, taken together with that CP term forms the likelihood ratio, which is a measure of how far your evidence, here what your degree of belief is and what your reliability is, confirm q. [How good an indicator are these things you believe about yourself of the truth about q?] The other terms are priors on that evidence that then give you the posterior probability of q. Both that measure of degree of confirmation in the bold black and red terms and this blue term are going to affect *how much* of a difference the new 2nd-order evidence will make to your degree of belief in q. The blue term is interesting since it says how confident you are in those beliefs about your belief and reliability. You ask yourself how reliable you are and you give that answer y. But you are not certain that it's y, you give it some confidence less than 1 because you evidence is always quite incomplete.

How much less depends on the quantity and quality of the evidence. A small set of data about your reliability is going to make the blue term closer to 50%, which will then mean the data doesn't affect your degree of belief in q much. This also addresses the problem that Rachael talked about yesterday, where once you see three cases your degree of belief in the fourth is determined in a way that seems wrong. Yes, there's a unique answer to what it is, but it's not determined by three data points.

It depends on what you judge your reliability to be, which you know three data points aren't a good indicator of, so if that's the only data you have there's almost no movement from your initial confidence.

Another issue about the quality of your evidence is *relevance*. You may be given by God a track record of all of your occasions of giving proofs of mathematical statements and you wonder how that should affect your confidence in your current proof. (Thanks Mike Caie for this example.) The first thing to say is that though God may be infallible about your track record, (how do you know it's not the devil talking in your ear and) that record is finite, and doesn't determine your calibration curve, which is a real valued function. But second, if it contains all of your proof-giving episodes it contains a wide variety of different mathematical statements as conclusions, and some that were much easier than the current one. How you did there doesn't make as much difference to how you do here as the cases where the proofs were just as hard; the easy ones will be largely probabilistically. If the track record contains harder proofs then they could be quite relevant as indicating a level of skill above the current case. Compare to the eyewitness testimony case where the question is exactly the same every time: Was that the perpetrator? This factor of relevance is automatically measured in the estimate of reliability given the track record that's going into the Jeffrey conditionalization.

But I think Teddy's concern may be that the calibration curve is *difficult* to know enough about to make a re-calibration not be distorting. For example q has to be a similar but *independent* question in the same subject matter. Rain each day is a nice example but how many cases are like that? In implementing we're using databases on classification problems – e.g. cancer or not on the basis of test results, where, like rain, it's easy to get streams of data about the calibration of the model. You just keep track of its performance. And, each data point is from an independent case. There are databases like this with 250,000 data points easily. When do we have this kind of information in daily life?

Well, we have statistics about the 120 million people who are married and what determines the couples' success or failure are independent (except when they're swapping, but that's going to be largely local, and lots of cheating is with the unmarried – also, some marriages dissolve without cheating). For our second case, the premise of the pessimistic induction is that there is a large group of independent yet similarly concocted and supported failed theories (although judging that track record for reliability is a weak spot in the pessimist's argument). And the person who should be worried about her confidence that the murderer is number 3 is looking at a mountain of empirical evidence about eyewitness testimony. The reason it worries us is that on all these independent occasions the same question – facial recognition – is being asked and answered by largely the same method – visual human memory – and you are a human being using visual memory for facial recognition.

And I think this sense of pressure for re-calibration that an ordinary person didn't feel before is coming from the fact that evidence about ourselves and our reliability is growing as never before.

Another thing to note is that the second-order update is less computationally intensive than re-training the model, at least the way we did it. It's computationally lightweight, which surprised me given that it's second order. The second order aspect was what got me worrying about tractability. [Oh, but it's probably lightweight because we're not doing it Bayesian, with priors. Still, the comparison in the Bayesian context may hold between Bayesian re-training and Bayesian re-calibration.]

SLIDE 57: So, now suppose you have all the evidence you could hope for. Another thing to consider is whether re-calibrating disturbs the bayesian's success in the long run. We know that 1^{st} order conditionalization converges to the true probability in the long run (under favorable evidence conditions). What happens if you have an infinite stream of evidence allowing you to apply Re-Cal indefinitely (in addition to first-order conditionalizing on the evidence? Here I'm changing to the chance version of the rule because for that kind the theorem I need is proved. This is a more explicit version of the rule.

READ Notice that you will end up with the right degree of belief in q if you believe the right chance function, i.e. calibration function. ….Whatever degree of belief you currently have in q, that function will tell you the true probability of q, and Re-Cal tells you to change to that. So the question is whether you can converge on the correct chance function, your own calibration function.

The evidence stream is ordered pairs of degrees of belief and rain or not, which can then be assembled into bins to get frequency of rain when you have a given degree of belief.

SLIDE 58: Jim Hawthorne proved a convergence theorem for Re-Cal using his likelihood ratio method, which I really like. Suppose you will be fed *separating evidence*, that is h, -h predict at least some different outcomes in the stream of evidence you're going to get. (Modulo Hume, brains in vats, conspiracies of nature, we have that.)

Then if h is the true hypothesis it is probable that you will see outcomes that rule out all the –h hypotheses within a certain number of trials. (Law of Large Numbers) In the infinite long run you definitely will, but he's also able to put a lower bound on *rate* of convergence based on how high the likelihood ratio is. (LR directly measures quality of the evidence. You actually flip it over to work with it because when the LR gets higher the denominator is going to zero.)

[Proviso: He proved it for two different degree of belief functions where I have to use 1. This is a problem because his proof is for strict conditionalization, and Re-Cal with one degree of belief function on itself and a strict conditionalization is going to be coherent there only when x=y, i.e. only when it's not necessary. There's only one way to go here for my project, which is to prove it for one function and in Jeffrey conditionalization. This is not yet done. However, there is reason to be optimistic (besides the fact that Jeffrey is a generalization of strict conditionalization). Skyrms (1980) showed that 1^{st}-order Jeffrey conditionalization is equivalent to a second-order strict conditionalization. So, 2^{nd}-order Jeffrey is equivalent to a third-order strict conditionalization.]

SLIDE 59: So what we have is that a LRCT theorem can be proven for the chance hypothesis (function) Of Re-Cal* [the calibration curve that is the second conjunct in the condition of Re-Cal] , i.e., where $h = Ch(q_v,P(q_v)=x)=y(q,v,c,x)$

(* because Re-Cal is written in chance, and because of the proviso of the previous paragraph.)

So, not distorting in the long run. Whatever satisfaction a convergence theorem gives us at the first order we still have if we add re-calibration.

SLIDE 60: Teddy has another important objection. As I said earlier was argued by Dawid and Teddy, and others, 1^{st} order conditionalization alone will make you subjectively calibrated in the long run, so why do re-calibration in the short run? This is a serious question I don't have a complete answer to. But the first thing to say is that Hawthorne's theorem shows more than *subjective* calibration in the long run. You converge to a confidence in q matching the *true* probability of q because you converge to your *actual* calibration curve which then allows you to always do a perfect *objective* re-calibration.

The other thing about his theorem is those lower limits on how long it takes to converge. This says something to the issue of whether Re-Cal improves your calibration at finite times: if the sequence is *long* enough then it definitely does. And if the length of the stream of evidence needed is realistic for us then a theorem of Schervish shows that what we can have: [How long does it have to be? If you can fill in the blanks in Jim's equation you can figure it out.]

SLIDE 61: If a forecaster is not well calibrated over a given (finite) sequence of events, then his well-calibrated counterpart *outperforms* him in similar decisions taken over this sequence. This means the calibrated one loses less utiles.

A simple case illustrates this. Suppose you have a person whose model bears no relation to what's producing the weather, so the confidences he comes up with about rain tomorrow have a random relationship to the facts. He has confidences like 80%, 65%, etc., but he's right half the time. He's betting his confidences -- usually more or less than 50% -- and so he's losing frequently. If he looked back at his track record as a set of ordered pairs, then he would see that of all the occasions when he said 80%, half of those were days it rained, and the same for every confidence. Now via Re-Cal he would move closer to 50% in his betting, and protecting himself against losses. He has no further ability to predict rain, but he's losing less money because he knows he doesn't know anything.

This case also illustrates that second-order evidence can be stronger than first-order, at a given finite stage. This guy was not in a position for the 1^{st} order evidence to be informative (yet), but the 2^{nd}-order evidence definitely was.

SLIDE 62: On this question of the added value of re-calibrating, there are some advantages that I see from a very general point of view.

First, if you think CP is a synchronic rationality constraint, then you should follow Re-Cal. It helps you do your best to stay in line with CP.

(Speaking of obligations, Re-Cal is just an instance of conditionalization, so if you endorse that as a learning rule then you owe me an argument why not to use this instance.)

Adherence to PP is not preserved under Jeffrey conditionalization. (Just proved this year by Nissan-Rozen.) I don't know if CP is preserved (though N-R's particular argument doesn't show it's not). However, Re-Cal gets you back to CP if you've fallen out so if you're always re-calibrating before settling on a confidence then you're okay vis-à-vis CP.

[Third, Re-Cal can coherently change extreme degrees of belief, to the other extreme or to non-extreme values in a principled way. I will come back to this.]

The bad predictive consequences of biasing assumptions in the model governing first-order conditionalization can be corrected by using Re-Cal, without knowing what those assumptions are, without re-training model. (We saw that with the random predictor above.)

In addition to what's on the slide, one of the things we're seeing is a computational advantage to re-calibration. Surprisingly, re-calibration, which we are doing in a roughly (but only roughly) Bayesian way by my equation, is computationally really cheap, cheaper than re-training the model. We're doing lots of comparisons to see added value, and see some conditions under which it's better to re-calibrate than to retrain, but nothing solid yet.

SLIDE 63: Suppose you are certain of q and read in a reputable journal that people who are certain are sometimes wrong. In a principled way you can update your confidence in q to something less than 1 because x need not equal y. It's an update rather than an exogenous change. That's for an empirical proposition.

With logical and mathematical proofs, the probability of a theorem is either 1 or zero, and that is so conditional on anything, so y in Re-Cal has to be the correct extreme value if you want to be coherent. However, x doesn't have to be extreme in Re-Cal, so it is describing what you should do if you find yourself (in fact incorrectly, incoherently) at some confidence in a logical/mathematical proposition, and your evidence is suggesting y is some other value. It tells you to move toward that value. And it seems that if your evidence is good you will eventually converge on that right confidence, 1 or 0, if your evidence stream is good. The firs-order Bayesian can only say that if you have the wrong confidence in a logical/math'l proposition then you are wrong. I can tell you how to use evidence to improve.

SLIDE 64: Another natural question is how far this analysis is restricted to a Bayesian framework. I think it both is and isn't. For my intuitive questions I think it's very important to use a system that puts a big emphasis on global consistency because how you could possibly be consistent while doubting yourself is prima facie a big part of the intuitive problem.

Secondly, recall that the updating being a conditionalization rule is how it escaped the complaints about hedging and informativeness. So that may be a challenge in a different formulation.

But finally I'd be happy if there were many ways of doing it and we could see the added value. The implementation we're doing isn't totally Bayesian though I hope to do that eventually.

SLIDE 65: To sum up the general points, I've argued

-- There's a connection between higher-order evidence questions and re-calibration.

and a connection between these and discussion of the Principal Principle (CP), which is not that surprising because calibration concerns the relation between subjective and objective probabilities.

-- A second-order, Bayesian, re-calibration rule can be formulated and defended by the independently appealing Conditional Principle.

I think the two biggest questions are:

-- Does it preserve coherence?

-- What is the added value?

We're supposing here that your creationist theory makes determinate predictions, so that we have ordered pairs <degree of belief, what happens when I have that degree of belief>.

SLIDE 51:

We can save ourselves from distorted predictions that come from assumptions that we have and and that we use to generate our confidence in q. Even though evidence is coming through that should be falsifying that assumption there can be a shielding assumption (e.g. God is sometimes testing us). However, that first-order shielding assumption will generally not work at the second-order, where the evidence you are using is your degrees of belief and how often q happens when you have those beliefs. So, you can re-calibrated away the distortion. This doesn't identify for you which of your assumptions is suspect, but it means you don't have to be able to identify them in order to correct for their bad effects.

SLIDE 55: Neither you nor the other guy has pixels that indicate tigers, but you should withdraw confidence in q and he should maintain confidence.

SLIDE 57: Another objection from teddy's 1980 paper. My understanding of the objection is that since one doesn't know one's true calibration curve one can't use it to update. And if one did know the curve one wouldn't need to update. So the recommendation to re-calibrate is pointless.

SLIDE 58: Both of these things are true, but it's possible to have *evidence* about our reliability, for example a track record of ordered pairs of confidences we had and frequencies of q being true at a given confidence.

SLIDE 58: And secondly this same objection could be made to first-order conditionalization. We need not be certain of our evidence either when we update. We use our best evidence to the extent that that's appropriate.

SLIDE 60:

- Also, it’s unclear to me what the subject, with two probability functions with different degrees of belief on q, suppose the subject is asked to bet on q. Which function “answers?” (This is why you need a tight bridge principle if you go with a typed theory – and that is how Skyrms originally formulated that principle.) No silly, P’ isn’t hers. She answers with the outside.