Uncertainty and Undermining
©Jim Pryor
Princeton & NYU

DRAFT 2 -- for Formal Epistemology Workshop in Austin, May 2005

1. THREE INFORMAL EPISTEMOLOGIES OF PERCEPTION

Let's consider three informal epistemologies of perception. The theories agree that we have perceptual justification; they disagree about how we get it.

Throughout our discussion, I'll suppose you to have a visual experience as of your hand. We want to know what it takes for you to be justified in believing that experience.

Let LOOKS-HAND be the hypothesis that it looks to you as if you have a hand. Let HAND be the hypothesis that you do have a hand. Let BAD be a skeptical hypothesis that entails that you'd have the same experience yet lack hands. E.g., BAD might be the hypothesis that you're merely hallucinating hands, or the hypothesis that you only have plastic hand-replicas, or the hypothesis that you're a handless brain in a vat being fed hand-like experiences.

Two of our epistemologies agree that, in order for subjects who look to have hands to be justified in believing HAND, they'll need some independent justification to believe that hypotheses like BAD don't obtain. These epistemologies think we ordinarily *have* that justification; they disagree about where it comes from.

One of them says it comes from broadly empirical considerations: e.g., the fact that your past experiences have been patterned in certain ways, and the best explanation of those patterns makes it likely that someone who looks to have a hand does. I'll call this the Inference-to-the-Best-Explanation (IBE) theory.

The other says your justification to believe you're not in a BAD case comes wholly from broadly *a priori* considerations. Putnam's *a priori* argument that he is not a brain in a vat falls under this heading. So too does Crispin Wright's theory, according to which we have *a priori* (but defeasible) entitlements to believe various BAD hypotheses don't obtain. I'll call this the A Priori Entitlement (APE) theory.

Our third epistemology denies that subjects need to have independent justification to believe that BAD doesn't obtain. It says that experiences as of hands are able to give some *prima facie* justification to believe HAND even *without* the help of independent justification to believe BAD is false. It's enough that you also lack justification to believe BAD is *true*. If you acquire justification to believe BAD is true, that will to some degree *undermine* the justification your experiences give you to believe HAND. (But it won't *wholly* undermine it, until you become certain you're in BAD.) I've defended this epistemology of perception elsewhere, and I call it Dogmatism.

If you're a Dogmatist, it's natural to go on to say that the prima facie justification your experiences give you to believe HAND also contributes, prima facie, to the justification of hypotheses entailed by HAND, e.g., that you're not a handless brain in a vat—despite the fact that, if you *were* a handless brain in a vat, you'd be having these very same experiences. I'll call those Dogmatist theories that go on to say this Moorean Dogmatism. (This is the only version of Dogmatism that I'll ask you to consider.)

The IBE theorist, the A Priori Entitlement theorist, and the Moorean Dogmatist largely agree about *who* is justified. The only examples they diverge over will be far-out ones (e.g., a reflective and conceptually sophisticated baby who hasn't yet had any perceptual experiences). For most ordinary cases, these epistemologists will agree that you *have* justification to believe not-BAD, and that you also have justification to believe HAND. What they disagree about is *the epistemic priority* of those two pieces of justification.

The IBE theorist and the APE theorist further disagree about *what your justification to believe not-BAD consists in*. For example, the IBE Theorist thinks that your past experiences make a positive contribution: your justification to believe not-BAD comes from the fact that those experiences were patterned in certain ways. The APE theorist, on the other hand, thinks that your past experiences have merely failed to step in, and defeat, some justification that was already in place without them.

Now, regardless of whether you think I'm backing the right horse, here, I hope you'll agree that this dispute is *intelligible*. It should be a substantive epistemological question which of these approaches is right, about any given subject matter. You might,

for instance, think that Moorean Dogmatism is wrong about perception; but that a structurally analogous view is right about introspection.

We ought to want our *formal* epistemologies to be able to represent these different options. If any formalism is unable to formulate the debate I described, I take that to be a weakness in the formalism. If any formalism makes the debate *unintelligible*, I take that to be a life-threatening weakness.

2. BRINGING IN THE BAYESIAN

Roger White[1] argues that Moorean Dogmatism violates Bayesianism, and so should be rejected. White voices a preference for something like the APE theory instead.

Let's consider some of White's complaints.

White argues that by Bayesian lights, prob(BAD|LOOKS-HAND) > prob(BAD). This is because BAD *entails* that it'll look to you as if you have a hand; which you weren't antecedently certain would happen. Now, it's not *transparent* how to move from the fact that you have a given experience to a decision about what hypothesis to update on. But it's standardly assumed that if you have an experience as of a hand, you should update on the hypothesis LOOKS-HAND. Hence, upon having an experience as of a hand, your new_prob(BAD) should = prob(BAD|LOOKS-HAND), and so should be > your old prob(BAD).

White claims that's a problem for the Dogmatist. For he interprets the Dogmatist to be saying that your new_prob(BAD) should be < your old prob(BAD). Why? Here's the reasoning he gives in §5 of his article. I've numbered the steps.

> [1] Dogmatism has the consequence that when it appears to me that there is a hand before me, I can gain justification, perhaps for the first time, for believing [not-BAD, i.e.,] that it is not a fake-hand, that I am not a handless brain-in-a-vat, and so on. [2] Now if I gain justification for a hypothesis, then my confidence in its truth should increase. [3] [The Bayesian

[1] "Problems for Dogmatism," forthcoming in Phil Studies; available online at http://www.nyu.edu/gsas/dept/philo/faculty/white/papers/dogmatism.pdf.

> complains:] But arguably when it appears to me that something is a hand, my confidence that [not-BAD, i.e., that] it is not a fake-hand should *decrease*. For since this is just what a fake-hand would look like, the degree to which I suspect it is a fake-hand should *increase*.

Later he repeats (I've substituted my nomenclature for his own):

> [4] [By Bayesian lights,] LOOKS-HAND raises the probability of HAND while lowering the probability of not-BAD... If my degrees of belief should conform to the probability relations outlined above, I should increase my confidence in HAND while decreasing my it in not-BAD. [5] And this is at odds with Dogmatism, which suggests that I might *gain* justification for not-BAD.

However, the Bayesian claims that White is marshalling aren't claims that the Dogmatist has denied, or needs to deny. The Dogmatist says that the justification your experiences give you to believe HAND *will contribute prima facie justification towards* not-BAD. That's not a claim about your all things considered credence in not-BAD. There may be several things going on epistemically at the same time. Simultaneously with justifying you in believing HAND, your experiences may also justify you in believing *you're having them*, and you might reasonably take that to make it somewhat *more* likely that you're in BAD than you antecedently thought. What the all things considered effect on your credences will be is unclear. The Dogmatist can allow that you end up more justified in believing you're in BAD than you started out. He just wants your experiences to tend to raise the probability of HAND *more* than they do the probability of BAD.

I agree then that the Dogmatist makes claims [1] and [5]. These claims have to do with "gaining justification," and in particular, with gaining prima facie justification. The Bayesian makes claims [3] and [4]. They have to do with relations between all things considered credences. It's not clear whether there's any conflict here, until we have a translation of the Dogmatist's claims into the Bayesian apparatus. White assumes one such translation (in [2]), and since the Dogmatist hasn't said much to guide him, it's rhetorically fair for him to do so. But as it turns out, I disavow that translation.

How then *should* we translate the Dogmatist's claims into the Bayesian apparatus? That's a question we'll be wrestling with.

Let's consider another of White's criticisms. He argues that prob(HAND|LOOKS-HAND)≤prob(not-BAD|LOOKS-HAND), and as we've already seen, the Bayesian will regard prob(not-BAD|LOOKS-HAND) as < prob(not-BAD). White writes:

> Hence, [6] prob(HAND|LOOKS-HAND) < prob(not-BAD). [7] So its appearing to me that this is a hand can render me justifiably confident that it is a hand, only if I am already confident that it is not a fake-hand.

The last claim [7] looks like something the Dogmatist would deny; but only when we interpret the "already" there as saying something about epistemic dependence. The Dogmatist denies that your justification to believe HAND *depends on epistemically prior* justification to believe not-BAD. But the Bayesian premise [6] doesn't obviously *say* anything about epistemic dependence or epistemic priority. It claims that a certain inequality holds between two all things considered credences. It's not clear how we *get* from claims of type [6] to claims of type [7].

The lesson that's emerging is that the Dogmatist and the Bayesian are using different vocabularies, and that it's hard to assess whether there's a conflict unless we have a translation manual. I think the translations White makes use of look reasonable at first glance, but I don't think they withstand scrutiny, and I don't think they're correct. *I* at least don't want to make the probabilistic claims that White interprets me to make.

I think the dialectic here is rather complicated. I'm going to be proposing a replacement for Bayesianism: a formal theory that's more hospitable to Dogmatism. But it's *not* because I agree that Bayesianism implies Dogmatism is false, and I want to hold onto Dogmatism at any cost. Rather, it's because I think that *even to properly express* the debate between the Dogmatist and his opponents, we need a theory that's expressively richer than Bayesianism.

What are the notions we need, if we're going to translate the debate between the Dogmatist and his opponents into formal terms?

1. Well, as we saw, we'll need some way to cash out the notion of "contributing prima facie support" to a hypothesis, without its necessarily being the case that that

hypothesis comes out, all things considered, ahead. It's obscure how to model that kind of contribution in Bayesian terms. Bayesianism is a theory about all-things-considered effects.

2. As we saw, the debate between the Dogmatist and his opponents concerned what your epistemic relation is to an *undermining* hypothesis, BAD. What does that consist in?

Justification can be defeated in different ways. Suppose my brother tells you that his landlord is shifty-looking. That gives you some justification to believe that his landlord is dishonest. One way for that justification to be defeated is for my brother's roommate to tell you that their landlord is *not* shifty-looking. That evidence *opposes* my brother's testimony. Its intuitive effect is to give you some justification to believe the opposite. Another way for your justification to be defeated is for you to learn that my brother's landlord is an active member of his church and donates generously to charity. This evidence *narrows the reference class*. Shifty-looking people are in general likely to be dishonest, but shifty-looking people who are active in their church and so on tend not to be. A third way for your justification to be defeated is for me to tell you that my brother never met his landlord, and just has a prejudice against him because of a disagreement over the rent. This evidence attests to my brother's *not being in a position to know* what his landlord looks like. So it *undermines* the justification my brother's testimony gave you to believe the landlord is dishonest. Intuitively, it doesn't give you any special reason to believe the landlord *is* or *looks* honest; he *may very well* look shifty and be dishonest. My testimony just gives you less reason to rely on my brother's word for it.

Hybrids of these different kinds of defeat are also possible. The BAD hypotheses we've been considering both oppose and undermine.

The undermining kind of defeat is intuitively distinctive, and it has played an important role in informal epistemology. But it's obscure how to capture it in Bayesian terms. *Informal* epistemologists usually gloss the notion like this: opposing evidence counts in favor of not-P; undermining evidence doesn't (or at least, needn't) do that, but instead it tells you that (some of) your existing evidence for P is unreliable. In a Bayesian

setting, though, that won't do. The Bayesian will regard *both* kinds of defeating evidence as increasing the proportion of your credence that should go to not-P, instead of P. So both can be construed as "counting in favor of not-P." Similarly, *both* kinds of defeating evidence should make you more suspicious of your earlier evidence for P: for, since it's now more likely that not-P, it's now more likely that your earlier evidence for P is evidence for a falsehood, and hence, more likely that there's some defect in the evidence.

It's not clear how to cash this distinction out, in Bayesian terms. I struggled to do it for a couple of years, and I read anything I could find, and digest, about how other people would do it. If we consider only the simplest cases, we can make headway using only Bayesian resources. But when we try to account for *hybrid* forms of defeat, too, I'm pessimistic whether a Bayesian can do it.

3. The Dogmatist and his opponents will agree that ordinary subjects have justification to believe *both* that HANDS and that they're not in BAD. But they disagree about whether their justification to believe HANDS *transmits to*, or whether it *presupposes*, justification to believe not-BAD.

There's a difference between the following kind of

WEAKER CLAIM: Having justification to believe P (to such-and-such a degree) *suffices for* you to have justification to believe Q (to such-and-such a degree).

All parties to our debate agree that whoever has justification to believe HAND also has justification to believe not-BAD. That falls short of the

STRONGER CLAIM: Your justification to believe P is *part of what makes you* justified in believing Q.

The difference between these claims plays a prominent role in the literature on "transmission-failure." E.g., Crispin Wright describes a case where you see what appears to be the scoring of a soccer goal, a cheering crowd, and so on. He allows that, if you're justified in believing (P) that a soccer goal was just scored, you will *also* be justified in believing (Q) that a soccer game is taking place, as opposed, to say, the filming of a movie scene. That is, having justification to believe P suffices for having justification to believe Q. But Wright denies that the justification you have for P "transmits to," or itself lends any credence to, the claim that Q. That is, what justification you have for Q can't

have *come from* your justification to believe P. It had to be in place independently. (Plausibly this will be because you needed some justification to believe Q as *a presupposition for* being justified, in the way you are, in believing P.)

The contrast between WEAKER CLAIM and STRONGER CLAIM is one that epistemologists have only recently started discussing, but it has deep intuitive support. It's obscure how to capture *it* in Bayesian terms, too. But we won't be able to fully express the debate between our three epistemologies of perception without it.

If we had a formal account of what undermining amounted to, then we might try to leverage that into an explanation of transmission. If there are any ways to *undermine* your justification to believe Q, without thereby *also* undermining (or at least contributing to the undermining of) P, that would show that the justification you have for P doesn't presuppose or derive from your justification to believe Q. After all, it's possible to take the latter away while leaving the former in place. So (when the underminers are absent), there'd appear to be nothing illegitimate or question-begging about letting the justification you have for P contribute to the justification of Q.

But, as I said, we don't have a formal account of what undermining amounts to.

The lesson I want you to take away from all this is that it's not clear how to translate certain informal notions, like:

- contributing prima facie justification towards a hypothesis
- undermining defeat
- relations of epistemic priority, question-beggingness, transmission

and so on, into a Bayesian framework. But those notions are required just to *state* the difference between the different theories of justification I started with. So we shouldn't think this is *just* the Dogmatist's problem. The differences between IBE theory, APE theory, and Moorean Dogmatism, are really invisible to the Bayesian, until we have those translations.

I'm going to propose a new formal theory—one that has a much easier time modeling these informal notions. I'll present this new theory in stages. The way I'll frame my presentation is as an attempt to formalize what the Moorean Dogmatist thinks is going on in perception. The formal theory we get will have the expressive capacity to

model *all three* of the informal epistemologies; that's what I want to advertise as its main strength. You won't need to be a Dogmatist, yourself, to appreciate the fact that this theory lets us *formally express* the difference between what the Dogmatist says is going on in perception, and what his different opponents say is going on.

So keep that dialectical set-up in mind. Our immediate quarry is a way to formalize the Dogmatist's informal epistemology of perception. But the strength of our result will be its expressive capacity; and you won't need to be a Dogmatist to appreciate that strength.

Let's review what the Dogmatist's informal epistemology says. It says: on the one hand, we have naive perceivers, who have no independent evidence that they're in a BAD situation. For such perceivers, the "full weight" of their experience should count in support of the external-word propositions, like HAND, that they represent. On the other hand, we have perceivers who have acquired some undermining evidence: some evidence that they are in a BAD situation. Perhaps a ticker tape seems to run across the bottom of their visual field, saying "You are a brain in a vat…" Of course, that shouldn't make it *certain* that they're brains in vats, but it would constitute some evidence. For such perceivers, the Dogmatist will say, their experiences as of hands count less in support of HAND. They should be just as confident that LOOKS-HAND is true; but they can't reasonably be as confident of HAND, on the basis of their experiences, as the naive perceivers can.

What do we need, to turn that informal picture into a formal theory?

First we'll need a way to *formally characterize* the difference between the naive perceiver and the better-informed, more skeptical perceiver. Second, we'll need to figure out how each subject is *supposed to update* in response to his experiences. And third, the account we give of updating should illuminate *the distinctive way in which* the subjects' perceptual evidence for HAND interacts with *undermining evidence* for BAD. That is, the difference between undermining evidence and other kinds of negative evidence should show up in our formalism.

3. REPRESENTING AGNOSTICISM WITH SUPERADDITIVITY

To start, then, how should we formally characterize the difference between a naive perceiver and his better-informed, more skeptical counterpart? The naive perceiver should be *agnostic* about whether he's in the BAD situation. Orthodox theories of probability have notorious difficulties giving a satisfying account of agnosticism. I'm going to follow a different approach. I'm *not* going to model agnosticism towards BAD as a state of having initial evidence telling to some degree for BAD, and to some degree against it. Instead, I'd like to formally distinguish agnostic subjects from *any* subject who's acquired real evidence for or against BAD.

Subjects who are agnostic towards P	different from	Subjects who have some evidence for P, and some evidence against it

One important difference is that the credences of the better-informed subject should be *more resilient* in the face of new evidence than the credences of the agnostic subject. Intuitively, a new piece of evidence about P should impress the neophyte more than it impresses the expert who's heard lots of evidence on both sides. We want a representation of that difference to show up somewhere in our formalism.

Orthodox theories of probability may be able to provide that. But it's not obvious or universally agreed how. I've played around with a few different approaches, some orthodox and some not. In the end I've found myself gravitating towards an unorthodox, SUPERADDITIVE theory of probability.

Put most simply, a superadditive theory is one that doesn't force a subject to divide all his credence between P and not-P:

[==== credence in P ======= || ======= credence in not-P ====].

Instead, it allows the subject to do this:

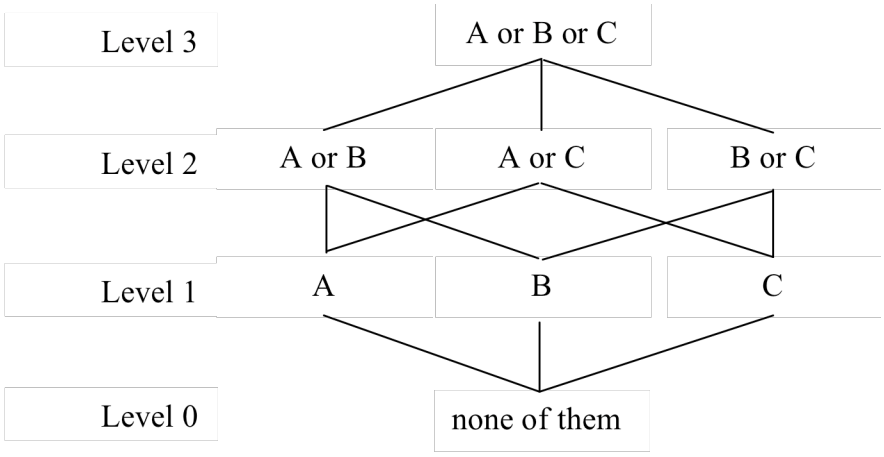
[confidence in P || ====== I dunno ====== || confidence in not-P].

The left-most quantity represents how much confidence you have that it's P that's true; the right-most, how much confidence you have that it's not-P; and the middle quantity

represents how much you remain doxastically "up in the air" on the issue. A completely agnostic subject would look like this:

[|| ========== I dunno ========== ||].

Here's how to think of such theories more generally. Let A, B, and C be mutually exclusive, exhaustive epistemic possibilities for you. Now consider a lattice built up out of these possibilities. It will look like this:



Henceforth, we'll ignore level 0. That leaves us with seven remaining nodes: A, B, C, A-or-B, A-or-C, B-or-C, and A-or-B-or-C.

Suppose that at each of these seven nodes there's a bucket. I give you one pound of "credence sand," and tell you to distribute your sand into the buckets in a way that models your confidence. The level 2 buckets are understood to "inherit" any sand you pour into their descendents: that is, any credence sand you commit to the A bucket is inherited by the A-or-B bucket, and also by the A-or-C bucket. Ditto for the level 3 bucket.

The Bayesian—and any probability theorist who follows the orthodox Kolmogorov axioms—will pour all of his sand into the buckets at A, and at B, and at C. If he's got some confidence in A-or-B, but he's unsure how much of it should go to A and how much should go to B, he'll just split the difference in some way.

A theorist who works with superadditive probabilities can pour sand into *any of the seven* buckets. If he's got some confidence in A-or-B, but has no reason to prefer one of those two hypotheses to the other, then he'll pour that much sand into the level 2 A-or-B bucket, rather than splitting it between the A bucket and the B bucket. This represents

how much he's "up in the air" between A and B. Of course, he might *also* pour some sand into the A bucket, and more sand into the B bucket. That sand will count as confidence in A-or-B, too. What the sand in the level 2 A-or-B bucket represents is his confidence *specifically* in A-or-B, that is, confidence in A-or-B that isn't committed any more specifically than that.

For example, someone who's completely agnostic between A and B and C, but certain that one of them is true, will pour all his sand into the level 3 bucket, A-or-B-or-C. Someone who regards it as 50% likely that C, and 50% likely that A-or-B, but who's not decided any further between A and B, can put half his sand in the C bucket and the other half in the A-or-B bucket.

Here's one terminology that's been used to describe these distributions. The amount of sand you have in the A-or-B bucket we call the "mass" that hypothesis A-or-B has for you. The mass of sand you have in the A-or-B bucket, *plus* the mass in all its descendents' buckets—the A bucket and the B bucket—we call the "credibility" that A-or-B has for you. The mass of sand you have in A-or-B, plus the mass of all its descendents, *plus* the mass of all the buckets *it* (or its descendents) descend from, we call the "plausibility" that A-or-B has for you. This last value is equivalent to 1 minus the credibility of *neither*-A-*nor*-B, that is, 1 minus the credibility of C.

Given a specification of how your beliefs assigns any one of: mass, credibility, and plausibility, we can derive the other two.

What corresponds most closely to our pre-theoretic thinking about "how confident you are in a hypothesis" is credibility. And on the theories we're now considering, the credibility of A-or-B can exceed the sum of the credibility of A and the credibility of B. It will do so whenever A-or-B has some mass of its own. So these views reject additivity:

$$\text{cred}(\text{A-or-B}) = \text{cred}(\text{A}) + \text{cred}(\text{B})$$

and replace it with superadditivity:

$$\text{cred}(\text{A-or-B}) \geq \text{cred}(\text{A}) + \text{cred}(\text{B}).$$

That's why they're called SUPERADDITIVE views.

On any superadditive view, the credibility of a hypothesis and the plausibility of that hypothesis can be understood as a lower and an upper estimate of that hypothesis' likelihood. So these views can be likened to approaches that work with interval-valued

probabilities, and do decision theory with interval-valued expected utilities. To take an example, if a some lottery ticket stands to pay $100, and you estimate the chances of its winning as: (credibility=40% … plausibility=80%), then you should be willing to buy the ticket if it costs less than $40, and willing to sell it if you can get more than $80. It's undetermined what you should do if when the ticket's price falls between $40 and $80.

On some interval-valued approaches to probability, the dynamics of belief are assumed to work just like the Bayesians say they do. We don't want to assume that here. The theory I'm presenting will employ a different dynamics. So this correspondence to familiar, interval-valued approaches should be expected only for the statics of belief.

We have to consider belief dynamics next. All we're assuming for the moment is this: Agnosticism corresponds to having all your sand high up in the lattice—perhaps as high as the wholly uncommitted bucket, A-or-B-or-C. As you acquire evidence, the lower buckets will start to fill up, and your confidence in the corresponding hypotheses will get more resilient. Resiliency will correspond roughly to the interval between credibility and plausibility.

4. DEMPSTER-SHAFER CONDITIONING AND UPDATING

OK, as I said, superadditivity is a feature of a *statics* of belief. It doesn't itself tell us how beliefs should rationally *change* in response to evidence. I want to explore one non-standard account of this. It's the account of updating associated with "Dempster-Shafer belief functions."

To present that account in the most intuitive way, we should dwell for a moment longer on a static issue, namely the issue of *how to condition your current beliefs on an assumed hypothesis*. Suppose that by your existing beliefs, A-or-B is not yet certain; but for some purpose you want to take A-or-B as given and reason downstream from there. Conditioning is the operation you use to do that.

We shouldn't assume from the outset that this conditioning operation is identical to the operation of updating on new evidence. Nominally at least, the operations are distinct. It needs some argument to establish any particular relation between them. I'm

calling the operation "conditioning," rather than "conditionalizing," to remind you not to assume yet that you know how it works, formally; or what relations it bears to the operation of updating on new evidence.

Let m(*x*) represent the mass your belief currently assigns to a hypothesis *x*; and m(*x*|A-or-B) represent the mass your belief assigns to *x* when conditioning on the assumption that A-or-B. If we had an account of m(-|A-or-B), then we might proceed to give the following account of *updating*. What happens when you *gain evidence* for a hypothesis, such as A-or-B, is that you acquire an extra quantity of sand. You distribute this new sand among your existing A-or-B buckets (that is, bucket A, bucket B, and bucket A-or-B), with each bucket *x* getting proportion m(*x*|A-or-B) of the new sand. Then of course you renormalize.

It's not usually presented like that, but that's in fact how the Dempster-Shafer account works.

The Dempster-Shafer account of updating just amounts to that intuitive picture, together with a proposal about *how much* extra sand you acquire, and a proposal about how to understand m(*x*|A-or-B).

What the Dempster-Shafer account says about *how much* extra sand you acquire is that it's proportional to the plausibility that the supported hypothesis (A-or-B) currently has for you. For a wholly agnostic subject, this will be 1. For a subject who's acquired some evidence *against* A-or-B, it will be less than 1. So the latter subject will get less sand to add to his A-or-B buckets. That's intuitively as it should be. Suppose you start out wholly agnostic, with 1 pound of sand in your top bucket. Then you acquire 0.2 additional pounds of sand in support of C. After renormalizing, your masses will be:

m(A) = 0
m(B) = 0
m(A-or-B) = 0
m(C) = 0.2/1.2 = 0.167
m(A-or-C) = 0
m(B-or-C) = 0
m(A-or-B-or-C) = 1.0/1.2 = 0.833

Next suppose you acquire some evidence that's intuitively *equally weighty* against C, that is, in support of A-or-B. If you added *another 0.2 pounds* of sand to your A-or-B buckets, that would tilt the scales *against* C. Intuitively, that's not what we want. We want the new evidence to *equalize* the lead that C has, not to exceed it. In order to equalize C's lead, we'd need to add exactly 0.167 pounds of sand to the A-or-B buckets, and then renormalize. This 0.167 = the original 0.2 you added to C * the current plausibility, 0.883, of A-or-B.

In general, then, the Dempster-Shafer account says that updating on evidence in favor of A-or-B works like this (for the moment, just regard k/(1-k) as an arbitrary fixed value):

new_m(A) = m(A) + k/(1-k) * plaus(A-or-B) * m(A | A-or-B)
new_m(B) = m(B) + k/(1-k) * plaus(A-or-B) * m(B | A-or-B)
new_m(A-or-B) = m(A-or-B) + k/(1-k) * plaus(A-or-B) * m(A-or-B | A-or-B)
new_m(C) = m(C)
new_m(A-or-C) = m(A-or-C)
new_m(B-or-C) = m(B-or-C)
new_m(A-or-B-or-C) = m(A-or-B-or-C)
(renormalized)

Suppose we apply that operation to a wholly agnostic belief-assignment, where m(A-or-B-or-C)=1 and all the other masses are 0. For a wholly agnostic belief-assignment, we can also assume that m(A-or-B|A-or-B)=1, and the other m(-|A-or-B)=0. That produces the result:

new_m(A) = 0
new_m(B) = 0
new_m(A-or-B) = k/(1-k) / (1+k/(1-k)) = k
new_m(C) = 0
new_m(A-or-C) = 0
new_m(B-or-C) = 0
new_m(A-or-B-or-C) = 1/ (1+k/(1-k)) = 1-k

What that tells us is that we can understand k as a measure of how much your evidence reasonably moves the belief of a wholly agnostic subject. His credibility should go from

0 to k. We can call k the "strength" of the evidence. Perhaps this value is the same for all possible kinds of evidence. Or perhaps it's different. I don't know. (For it to be different would mean that some kinds of evidence are intrinsically more probative than other kinds.) All I'll assume is that *a given* piece of evidence will have the same k-value or "strength" for every subject. I count two subjects as "acquiring the same evidence" only when that evidence testifies in favor of the same hypothesis, and has the same strength.

Of course, that doesn't mean that the *net*, *all things considered* effect of the evidence will be the same for different subjects. If subjects start with different background beliefs, then of course the evidence will have different net effects on their belief. But *something* needs to be invariant, for it to count as the same evidence. On the current account, what needs to be invariant is the evidence's "strength"—how much it should move the belief of a wholly agnostic subject—and what hypothesis it directly testifies in favor of.

We're still waiting on an account of how to understand m(-|A-or-B). Here's one intuitive proposal for how to condition your belief, on a superadditive theory. Consider the mass your existing belief system assigns to A-or-C. If you're assuming for the sake of argument that A-or-B, that is, that C is false, then the confidence you have suspended between A and C should presumably fall to A's bucket. Similarly for the confidence you have suspended between B and C; that should fall to B's bucket. Similarly for confidence you have suspended between A and B and C; that should fall to A-or-B. Any confidence you had committed to C you should abandon. Finally, you renormalize what's left. In other words:

$$m(A \mid A\text{-or-}B) = m(A) + m(A\text{-or-}C)$$

$$m(B \mid A\text{-or-}B) = m(B) + m(B\text{-or-}C)$$

$$m(A\text{-or-}B \mid A\text{-or-}B) = \underline{m(A\text{-or-}B) + m(A\text{-or-}B\text{-or-}C)}$$

(renormalized)

That's not the only procedure you might think of for conditioning belief, but it is a natural and intuitive one. If we combine that with what came before, we have the complete Dempster-Shafer account of conditioning and updating.

Let's consider two other ways you might think of to understand the operation of conditioning.

Proposal 2: Instead of your existing *masses*, let's instead consider *the evidence that gave you* those masses. If some of the evidence that produced your existing belief originally told in favor of A-or-C, then, since you're now *assuming* C is false, you should within the context of that assumption regard that evidence as informing you that A. Similarly for evidence that originally told in favor of B-or-C. Any evidence that told in favor of C must, within the context of your assumption that C is false, have *something* wrong with it, so you just ignore it. Proposal 2 says that the beliefs you *should* have, conditional on the assumption that A-or-B, are the beliefs that *would be* rational if you had re-interpreted all your existing evidence in that way.

Proposal 3: Instead of all that complicated re-interpretation, what if we instead assume that you've *now* acquired absolutely certain evidence for A-or-B. Look at how you'd update if that were so. The (unconditional) belief assignment that *would* result if you got certain evidence for A-or-B gives you the *conditional* beliefs you should *now* have, relative to the assumption that A-or-B. (This is the identification of conditioning and updating that we were earlier hesitating about.)

Each of these three proposals for how to understand conditioning has some intuitive support. It would be hard to choose between them. It'd be nice if we didn't have to choose. On some dynamics for superadditive belief, we would have to choose. But a beautiful virtue of the Dempster-Shafer dynamics is that we don't. On the Dempster-Shafer account of conditioning and updating, these three procedures all yield the same results.

I don't know whether the Dempster-Shafer dynamics are the only serious dynamics for superadditive belief to have that virtue. But none of the other theories I've explored have had it.

5. UNDERMINING AND BIASED EVIDENCE

So far so good. But we don't yet have an account of how undermining evidence works. To get that, we'll have to adjust the belief dynamics the Dempster-Shafer theory gives us.

Recall the informal picture of perceptual justification that the Dogmatists offer us. They say that for naive perceivers, who have no positive reason to think they're in a BAD situation, the "full weight" of their experiences should count in support of the external-word propositions (like HAND) that they represent. Perceivers who have acquired some undermining evidence, on the other hand, should be just as confident that LOOKS-HAND is true, but less confident in HAND.

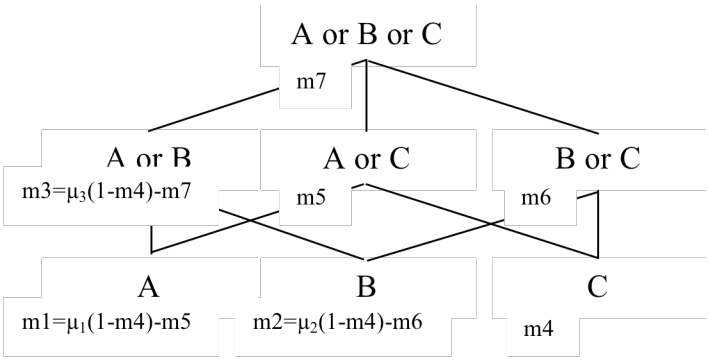
I'll say that your evidence is BIASED towards a hypothesis A when it testifies in favor of A, but there are other, not-A possibilities, like BAD, whose obtaining would entail that you have the same piece of evidence. Those other possibilities count as *undermining hypotheses* for that evidence.

If your evidence isn't at all biased, then neither will it be underminable. I'll call evidence of this sort "neutral." Neutral evidence may still be *fallible* evidence: it can be outweighed by a greater quantity of evidence for a competing hypothesis. But neutral evidence will be immune to the distinctive *kind* of *undermining* defeat we've been focusing on. For neutral evidence, the hypothesis that *you have* the evidence entails that *the hypothesis the evidence supports* is true. So there can't be any BAD hypotheses that give you the first without the second. The only room for error or uncertainty is about whether you have the evidence, in the first place.

As I understand the debate between Dogmatism and the other epistemologies of perception, it's a debate about whether the evidence our experiences give us is biased or neutral. The IBE theorist and the APE theorist take it to be neutral. When you have a visual experience of a hand, the only hypothesis that (directly) gets supported is LOOKS-HAND. Our new beliefs will come out in favor of HAND only if there's a prejudice already built into our existing, background, beliefs. (It may be there for wholly a priori reasons, or it may not.) On these theories, that's the only way for our beliefs to reasonably come out in favor of HAND rather than BAD.

The Dogmatist, on the other hand, takes experiences to support hypotheses that *aren't* just entailed by your having those experiences. So it's possible to have the experience without the supported hypothesis being true. That's just the sort of possibility represented by BAD.

To get a formal theory out of the Dogmatist's informal picture of perception, let's consider how the Dogmatist thinks our experience's epistemic effects should *differ from* the effects of some hypothetical neutral evidence, of the same strength. We'll consider a subject whose belief assigns the following masses:



That looks more complicated than it really is. I've just set up the variables so that the subject's m(A|A-or-B) will = μ_1, m(B|A-or-B) will = μ_2, and m(A-or-B|A-or-B) will = μ_3. We'll consider several different choices of μ_1, μ_3, and μ_3. Additionally, note that plaus(A-or-B)=1-m4. We'll hold this value fixed throughout.

We want to compare evidence that's biased towards A, but underminable by B, to neutral evidence of the same strength for A-or-B. Let's start with the neutral evidence. If it has strength k, then the Dempster-Shafer dynamics tell us to update as follows. Let X = k/(1-k)*plaus(A-or-B) = k/(1-k)*(1-m4). Then our new belief should assign these masses:

new_m(A) = m1+Xμ_1
new_m(B) = m2+Xμ_2
new_m(A-or-B) = m3+Xμ_3
new_m(C) = m4
new_m(A-or-C) = m5
new_m(B-or-C) = m6
new_m(A-or-B-or-C) = m7
(renormalized)

How should the operation differ when we update on the biased evidence?

We can motivate the following constraints.

1. Our account of biased updating should agree with the account of neutral updating about how confident to end up that A-or-B. For example, a Dogmatist should agree with the other epistemologists as to how confident we are that LOOKS-HAND is true. What this constraint comes down to is that, when you're updating on the biased evidence, you should get the same amount of new sand, X, to distribute to your A-or-B buckets.

2. When you're updating on the biased evidence, you should ordinarily assign *more* of the new sand to A than you do when updating on neutral evidence for A-or-B. That's how we model the bias.

3. Consider the case where μ_2=m(B|A-or-B)=0. That is, you have no positive evidence in favor of B, even when conditioning on the assumption that A-or-B. The dogmatist's picture is that undermining hypotheses should have no impact until you get evidence for them. So in this case, the Dogmatist will claim that the "full weight" of your evidence should go to A. That is, in this case you should acquire just as much confidence in A as you do in A-or-B. You should assign *all* your new sand to A. You should however be prepared to back off from A as evidence for B arises.

4. In cases where μ_2>0, you should assign less of the new sand to A. In those cases, the evidence's bias towards A is to some degree undermined by your justification to believe B. Still, *some* degree of bias should remain, so long as you're not yet certain you're in B.

5. To make the proposal commutative, it shouldn't matter whether your justification to believe B comes *before* or *after* you acquire biased evidence for A. If it comes later, you'll have to go back and *readjust* how you distributed any sand that's biased towards A.

6. Should the bias your evidence has towards A be allowed to count towards the justification of all of A's consequences? In particular, should it count towards the credibility of not-B (= A-or-C)? Can the bias your evidence has *towards* A be allowed to *discredit* B? I think we feel pulled in different directions about this.

On the one side, A and B are incompatible. So if the evidence *really is* going to be biased *towards* A, we think, it must to the same extent be biased *against* B. Any justificatory contribution to A should *constitute* a justificatory contribution to the

credibility of A-or-C, that is, to not-B. (It's compatible with that that there also be *other* effects on not-B's credibility, and that the *net* result is for it to go down.)

On the other side, we have some real *reluctance* to allow evidence that's underminable by B, evidence that you'd exactly *expect* to have if B were true, to justify you in *lowering* your estimate of B's likelihood. How could it do that? Philosophers often complain, "You'd be having exactly the same evidence *if you were* a BIV. So how can that evidence give you reason to think you're not?"

A virtue of a superadditive framework is that it lets us respect both pulls. By raising the mass of A, the biased evidence does contribute to the credibility of A-or-C, in other words, it does contribute to the credibility of not-B. But since the credibility of not-B + the credibility of B need not sum to 1, raising the credibility of not-B *doesn't mean* you have to lower the credibility of B. You can make the credibility of not-B go up without the credibility of B going down.

The dynamics I propose will work that way. Biased evidence won't lower the credibility of hypotheses it's subject to being undermined by. What that comes down to is that, when updating on the biased evidence, you should assign exactly as much of your new sand to B, as you would when updating on the neutral evidence for A-or-B. You should just tend to assign *more* of the sand that's left to A.

As we've seen, in the case where $\mu_2=0$, and both updating procedures give amount 0 of new sand to B, the biased updater should give *all* of the remaining new sand to A. The simplest way to accommodate the constraints I listed is to make that true in general. That is, the biased updater should *always* put as much new sand as he can in A's bucket, consistent with satisfying the above constraints. So, when your evidence is biased towards A, but underminable by B, and X=k/(1-k)*plaus(A-or-B), you should update like this:

new_m(A) =	$m1+X(\mu_1+\mu_3)$
new_m(B) =	$m2+X\mu_2$
new_m(A-or-B) =	m3
new_m(C) =	m4
new_m(A-or-C) =	m5

new_m(B-or-C) = m6

new_m(A-or-B-or-C) = m7

(renormalized)

A virtue of this proposal is that it allows us to keep the accounts of conditioning we gave in the previous section, and get the result that when the biased updater conditions on either of the assumptions:

- not-A
- not-B

he gets the same results that the neutral updater would get. Conditional on the assumption that you're not perceiving, the full weight of your experience counts towards your being in the BAD situation. Conditional on the assumption that you're not in BAD, it counts towards your perceiving. Both updating procedures yield that result.

(For simplicity, I've been pretending throughout, and will continue to pretend, that there's only one possibility in which you're perceiving and only one BAD possibility.)

That, then, is my proposal for how to update on biased evidence—and thus, my proposal about how evidence that's biased towards A interacts with evidence for hypotheses that undermine it.

6. CONCLUSION

I said before that the strength of this system is not just its ability to accommodate the Dogmatist, but rather its ability to accommodate *all three* of the epistemologies we considered.

The IBE theorist and the APE theorist will deny that our experiences give us biased evidence; and thus, in my terms, they'll deny that the justification our experiences (directly) give us is really underminable. Rather, they'll say, there are prejudices in our priors. We don't really start out in the position of a true agnostic.

The simplest model of that would be if we started out with neutral a priori evidence to believe we're not in BAD. On that model, the reasonable way to assign our belief, prior to having any experiences, would be something like this:

m(HAND) =	0
m(BAD) =	0
m(HAND-or-BAD, i.e., LOOKS-HAND) =	0
m(not-LOOKS-HAND) =	0
m(not-BAD) =	α
m(not-HAND) =	0
topmost mass =	$1-\alpha$

That would be the simplest model of an APE theory. Note that on this model, our a priori evidence for not-BAD isn't underminable. It might be *outweighed* by empirical evidence that we are in BAD, but it can't be undermined. We'll consider a more complex model in a moment.

The IBE theorist rejects the APE theorist's claim that, prior to having any experiences, we have more *un*conditional justification to believe we're in not-BAD than that we're in BAD. Instead, the IBE theorist maintains, that will only be true *conditional on* our having certain courses of experience. All we're unconditionally justified in believing, a priori, is something like this. Let ORDER be the hypothesis that we (will) have orderly courses of experience, that are best explained by our not being in BAD:

m(HAND & ORDER) =	0
m(BAD & ORDER) =	0
m(not-BAD & ORDER) =	0
…	
m(not-ORDER) =	0
m(not-BAD) =	0
m(not-BAD-*or*-not-ORDER, i.e., ORDER $\supset$not-BAD) =	α
topmost mass =	$1-\alpha$

On this view, your a priori credibility(not-BAD) = 0, but your a priori credibility(not-BAD|ORDER) turns out to be α. It's only by acquiring *empirical* evidence that you're in

ORDER that you become entitled to discount BAD. If your course of experience *weren't* orderly, you *wouldn't* be justified in believing not-BAD.

Of course, the APE theorist wants to *agree* that if your course of experience weren't orderly, you'd be less justified in believing not-BAD. But he wants to keep his account from collapsing into the IBE theorist's. In the basic case, he wants your justification to believe not-BAD to be a priori. It shouldn't need to draw support from premises about patterns in your experience.

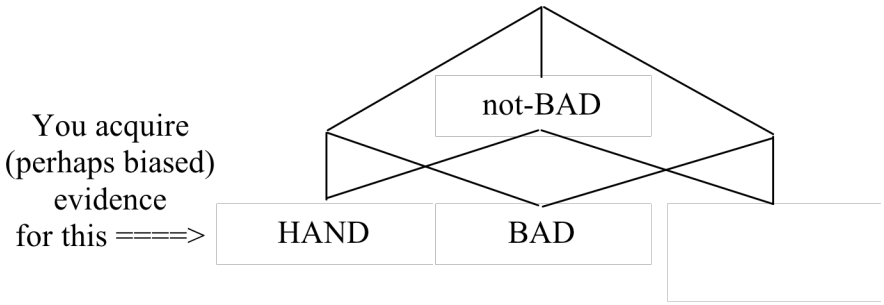
We could stick with the simple APE model I described a moment ago. Then you'd just have some quantity of a priori justification to believe not-BAD, which might be outweighed by contrary empirical evidence for believing BAD. A more sophisticated model would say that your a priori justification to believe not-BAD is *underminable*. It's *biased* towards BAD's being false, but certain kinds of disorderly empirical evidence can undermine that bias.

If that's the best way for the APE theorist to go, then he'll need an account of bias and undermining, just as much as the Dogmatist does. The difference is that the Dogmatist makes *our experiences* biased, and underminable, whereas the APE theorist makes our experiences neutral, but our a priori entitlement to believe we're not in BAD be underminable.

How does the theory I'm proposing represent the disagreement about *the relative priority* of your justification to believe not-BAD and your justification to believe HAND?

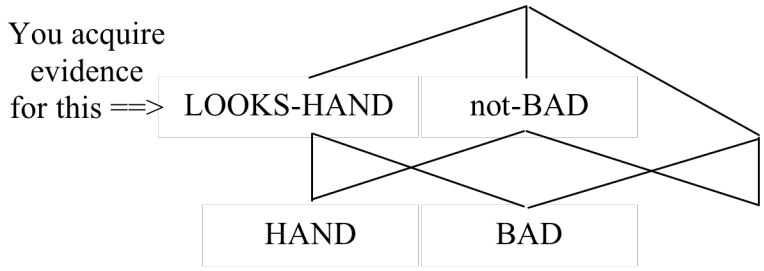
Easy.

A situation where your justification to believe not-BAD "comes from" your justification to believe the more specific HAND looks like this:



Since the credibility of not-BAD inherits all of HAND's mass, the net effect of your evidence will be to raise the credibility of not-BAD. (It need not *also* lower the credibility of BAD. Whether it does so will depend on whether the evidence is underminable by BAD.)

A situation where your justification to believe HAND "comes from" neutral evidence for the more general LOOKS-HAND, interacting with your background justification to believe not-BAD, looks like this:



If the mass of not-BAD is > 0, then your evidence for LOOKS-HAND will have the net effect of also raising HAND's mass. That's already predicted by the Dempster-Shafer dynamics for neutral updating.

I hope that conveys some of the flexibility, and expressive power, of my formal system. It enables us to express *each* of the competing hypotheses about the epistemology of perception, and to model the ways in which they differ. Perhaps Bayesians will be able to mimic all these epistemic dynamics. It's obscure to me how; but I won't say they can't do it. I am confident that they can't do it anywhere near as simply and elegantly.

I want to stress in closing that the formal system I've proposed doesn't help us *choose between* the different epistemologies of perception. As it should not. That's a choice we should make by doing epistemology, not by doing math.