Reconsidering Authority

Scientific Expertise, Bounded Rationality, and Epistemic Backtracking

Michael Strevens

To appear in *Oxford Studies in Epistemology*, vol. III

Abstract

How to regard the weight we give to a proposition on the grounds of its being endorsed by an authority? I examine this question as it is raised within the epistemology of science, and I argue that "authority-based weight" should receive special handling, for the following reason. Our assessments of other scientists' competence or authority are nearly always provisional, in the sense that to save time and money, they are not made nearly as carefully as they could be—indeed, they are typically made on the basis of only a small portion of the available evidence. Consequently, we need to represent the authority-based elements of our epistemic attitudes in such a way as to allow the later revision of those elements, in case we decide in the light of new priorities that a more conscientious assessment is warranted. I look to the literature in confirmation theory, statistics, and economics for a semiformal model of this revision process, and make a particular proposal of my own. The discussion also casts some light on the question of why certain aspects of science's epistemic state are not made public.

A truly individualistic epistemology would be a disaster—not on a desert island, perhaps, but disastrous for us, living in this world in which almost all information about anything important is mediated by other humans.

The question of the circumstances in which we trust other people's assertions—of the circumstances in which we attribute to them epistemic authority on the matters in question—is a familiar one. Also familiar is the question of the justification for knowledge acquired in this way. Less familiar, though perhaps due for revival, is the question of how to quantify such authority when it is attributed, that is of how much weight to give a proposition on the grounds that it has been endorsed by such and such an authority.[1] In this paper, I will ask a fourth question: once the degree of weight due to authoritative endorsement is determined, what do we do with it? How do we integrate this quantity into our epistemic outlook?

I will argue that there is good reason not simply to merge it with information derived from other sources; rather, we should give it special treatment, keeping track of it even after we have extracted whatever epistemic goodness it contains. The reason for this bookkeeping is the need—in the light of certain deficiencies characteristic especially, if not exclusively, of authority-based belief—to implement what I call *epistemic backtracking*: the need in certain circumstances to go back and adjust our original assessment of the authority, and to adjust accordingly all beliefs formed on the basis of that authority. I will search for a semiformal model of epistemic backtracking, a model formal enough to be integrated into, say, a Bayesian epistemology.

The theater of my investigation will be scientific inquiry, where epistemic authority is indispensable and where the question of what gives epistemic support to what and with what weight is never far from the surface.

1. On the importance of authority see Hardwig (1985); on the rationale for yielding to the epistemic influence of authority see Hardwig (1985); Foley (1994); Coady (1995); Goldman (1999); Kelly (2005); Lackey and Sosa (2006); Christensen (2007); on the question of what weight to give authority, see Lehrer and Wagner (1981); Kitcher (1993); Kelly (forthcoming), among many others.

1. Appealing to Authority

To a first approximation, nobody knows any science. The expertise of even a professional scientist constitutes only a speck in the vast constellation of scientific knowledge. What brings these isolated points of light together to form science's grand blueprint of the world is a social network, in which the principal epistemic material is scientific authority. Our acceptance of the scientific image of the world is, then, based on trust in what scientists say, whether we are civilians or scientists ourselves (Hardwig 1985; Hull 1988; Kitcher 1993; Shapin 1994; Goldman 1999).

From a scientist's perspective, the role of authority can be divided, roughly, into two parts: the proximal, that is, authority concerning matters directly relevant to the scientist's own field of study, and the distal, that is, authority concerning the rest of science.

Distal authority is familiar to non-scientists. It is the kind of authority that is attached, paradigmatically, to all-evidence-considered judgments about theories; in particular, it is the authority that stands behind the public acceptance or rejection of theories: "The caloric theory of heat turned out to be wrong," "There is at this time no credible alternative to the theory of evolution," "The jury is still out on dark matter," and so on. It is also the kind of authority that is more salient in the construction of the big scientific picture, in the big sense of *big*.

Reliable distal authority would not be possible, however, without proximal authority, the kind of authority that drives the short-range informational conveyor belts deep inside the factory of science, and that as such is of more immediate concern to the factory hands, the working scientists. It will be my focus in this paper.

Proximal authority's essential role consists in its being attached to propositions concerning three kinds of states of affairs: experimental outcomes, the initial plausibility of hypotheses, and the correctness of auxiliary hypotheses. Let me say something about each in turn.

Experimental Outcomes Can you rely on somebody else's experimental data? The live question here does not so much concern the experimenter's veracity—scientific fraud is rare (Hull 1988)—as their competence. No one normally doubts that the needle pointed where it was said to point, or that the rat went where it was said to go, or that the survey data came out in just the way reported in the journals. What is at issue is whether this correctly reflects the patterns of phenomena under investigation rather than constituting a mere experimental artifact—a smudge on the glass, an unexpected power surge, an unforeseen confounding variable. Some scientists have better reputations for producing reliable, artifact-free experimental data than others. Their results, as a consequence, have greater authority. (It is for this reason that *authority* seems a better rubric for the discussion than *testimony*.)

A particularly dramatic and unusually cross-disciplinary example is Pons and Fleischmann's 1989 announcement of the observation of tabletop cold fusion. Could these experimenters be trusted? Nuclear physicists investigating fusion by quite different means, knowing very little about the science of electrochemistry, had to put this question quite explicitly to their colleagues in other departments (Kitcher 1993, §8.2). The extensive and expensive effort to reproduce Pons and Fleischmann's results was undertaken only once it emerged that the researchers' work had in the past indeed been very reliable, a track record that endowed their announcement with considerable authority. In more quotidian science, the same sorts of decisions about what to take seriously, what to attempt to replicate, and so on, are taken on similar grounds every day.

Plausibility of Hypotheses Einstein is overthrown and quantum physics reformulated about once a month in my in-box. My friends in physics learn of these scientific revolutions in the making even more frequently still. We resist the temptation to investigate the workings of the radical new theories brought in with the morning email, however, because we do not trust their sources.

This is an extreme case. But it is a fact of scientific life that not every hypothesis proposed by a credentialed scientist can be tested exhaustively at the same time. Priorities must be established, and these priorities will depend to a certain degree on the credibility of the author. This is not to say that voices from the margins will be snuffed out; they will, however, have to work to be heard.

Scientific authority enhances not only the plausibility of hypotheses—our confidence in their truth—but also our confidence in their fruitfulness. We judge a hypothesis more likely to lead to valuable applications, to other interesting hypotheses, perhaps even to profound theoretical innovation, if it comes from someone with the right reputation.

Auxiliary Hypotheses Even if an experiment tells you what it purports to tell you about a given phenomenon—that a tabletop fusion apparatus produced more energy than it consumed, or that the speed of light from the sun is the same in several different inertial frames—its bearing on a theory is typically, perhaps inevitably, mediated by other hypotheses that are not a proper part of that theory, often called (when they serve in this capacity) auxiliary hypotheses.

Neither experts on the experiment nor experts on the theory may be experts on all the auxiliaries. Help must be sought, and that means reliance on authorities from neighboring fields. This sort of epistemic aid is both distal and proximal: distal in its sources (workers in other specialties), proximal in its application (the interpretation of experimental data in your own specialty, including your own data). For its proximal relevance, it will be included in the discussions to follow.

Some remarks. First, for simplicity's sake, I have talked only about experimental science, as opposed to science based largely on observations of phenomena not generated by experiment, such as cosmology or paleontology. But what I say about the role of authority in determining the reliability of

experiment goes also for these other kinds of empirical observation.

Second, the most striking immediate effects of variation in authority are practical: they concern which experiments to take seriously, which experiments to attempt to replicate, which hypotheses to test. But the decisions in question draw on purely epistemic consequences of authority: confidence that an experiment was not marred by artifacts, or that a hypothesis is well-conceived, motivated and potentially fruitful. It is of course these epistemic ramifications of scientific authority that are my official topic.

Third, by focusing on the uses of authority concerning proximal rather than distal matters, I have largely put to one side deference to authority that is based on the authority's being acquainted with a wider range of experimental evidence than the deferrer (though deference concerning auxiliary hypotheses will to some extent keep this aspect of scientific authority in play). I am interested in the way that authorities differentially confer weight on what is published in the scientific literature, rather than their use as a substitute for a review of that literature, convenient though they often are in that respect.

2. Quantifying Authority

How do scientists assess one another's reliability, and thus authority, with respect to experiment, hypothesis, and so on? And how do these assessments feed into the epistemology of science more generally? It is the second question that is my topic here, but I will motivate my investigation with some remarks about the first.

Let me begin by constructing a makeshift epistemic framework for the quantification of scientific authority. I will work within a probabilistic epistemology, in which all propositions or putative states of affairs are assigned an epistemic probability that changes as empirical evidence and other information—such as endorsement by scientific authority—come in. (No particular interpretation of this probability is supposed. It is intended to be

compatible with, but not only with, the Bayesian epistemology of science.)

Authority, I will suppose, is a scalar property: it comes in degrees, though degrees of what I will not need to say. As a placeholder for this unknown epistemic stuff I will sometimes use the term *weight*. The greater the authority, the greater the weight that comes with endorsement by that authority, and the greater the weight that comes with endorsement, the more the probability of the endorsed proposition stands to rise. (Obviously authority is topic-relative: a particular scientist will have different degrees of authority with respect to different propositions.)

How, then, to assess authority-based weight? There is general agreement that weight assessment ought to quantify the reliability of a scientific authority's judgments (Polanyi 1958; Kitcher 1993; Goldman 1999). The problem is what method to use to ascertain reliability.

Kitcher's solution is *calibration*, a kind of straight induction (Kitcher 1993, §§4–7). Calibration comes in two flavors, direct and indirect. In the direct case, the assessor proportions authority-based weight based on their own observation of the accuracy of the putative authority's judgments. In other words, they follow the literature closely, tracking a scientist's pronouncements and the ultimate fate of those pronouncements. An experimenter whose results stand up under scrutiny is attributed high authority as a producer of empirical results in their field; a theorist whose hypotheses generate interesting empirical work, discussion, further hypotheses, or perhaps even turn out to be correct is attributed high authority as a proposer of hypotheses in their field.

No one can directly calibrate the authority of more than a few scientists at a time, so there is a need for indirect calibration, in which a scientist adopts the direct calibrations of other scientists, as when fusion physicists sought out electrochemists to ask about Pons and Fleischmann's experimental credentials. With indirect calibration comes the problem of aggregating any differences in the estimates of reliability. If one expert tells you you can trust

so-and-so, but another is not so sure, what to believe?

The natural calibrationist response is second-order calibration: assess the authorities' skill as calibrators. Or find an authoritative second-order calibrator and use *their* assessments of first-order calibrational ability to settle the question. There is then a further question of what to do with these second-order judgments. Find the first-order calibrator with the greatest authority and use their judgments alone? Or use a mix of the different first-order opinions, weighted by second-order assessments of their reliability? The former strategy has the appearance of a quick and dirty heuristic, and indeed, the consensus seems to lie with the latter approach (Lehrer and Wagner 1981; Kitcher 1993). The same considerations suggest that, even when you have directly calibrated some scientist's authority yourself, you ought to mix your own assessment with that of others, in effect treating yourself as just one first-order calibrator among many. (For a dissent, see Kelly 2005.)[2]

Kitcher's method of calibration is a reasonable first approximation to the way in which scientists go about the task of assessing authority. It is not necessary, however, that such assessments be driven by straight induction. There is room for what a Bayesian would call prior probabilities: attributing greater authority, say, to a scientist who had a prestigious advisor or who has a prestigious academic position than to someone more obscure, on the grounds that there is some correlation between these factors and a scientist's reliability. Such tentative estimates of authority will of course diminish in importance as information about the scientist's actual success rate comes in. (Kitcher classifies the authority bestowed on scientists in virtue of pedigree and so on as "unearned authority," which suggests that it is allocated quite independently of the calibrational endeavor—an interesting and anthropologically not implausible alternative to my tentative proposal that prestige stands proxy for reliability.)

2. There is a substantial literature in statistics on calibration and related techniques; see for example Schervish (1989).

There is much more that could be said about techniques for calibration and other methods of authority assessment. Indeed, there is much more that has been said, both concerning the internal rationality of such procedures, and concerning their social-epistemological utility. I will go no deeper into the topic, however. I have what I need already, namely, some relatively uncontroversial, qualitative observations about the process of authority assessment: (a) that authority-based weight is based on an estimate of a certain kind of reliability; (b) that any conscientious attempt to assess this reliability—any attempt at calibration—will have both direct and indirect components; (c) that indirect calibration requires second-order calibration, that is, estimates of scientists' ability as first-order calibrators; and less importantly (d) that proxies for reliability, such as a prestigious publication venue or academic position, may figure in the calculation.

3. Questioning Authority

The qualitative observations made in the previous section show that authority, even when understood as a quantification of reliability, is a rather slippery thing to measure. Consequently, I will argue, we should regard judgments of authority—and more generally, assessments of net authority-based weight, which I will sometimes for brevity's sake also call authority assessments—with a grain of epistemological salt, as works in progress rather than as final judgments. Of course, any epistemic probability is a work in progress in some sense: as our state of knowledge changes, epistemic probabilities will change. But judgments of authority are provisional in a more profound way: they are, I will demonstrate, apt to change even when our state of knowledge stays the same.

I should say right away that I am not advocating a deep skepticism about assessments of authority. Scientists' judgments of authority, though in a certain sense flawed, do an acceptable job of tracking what they are supposed

to track—roughly, reliability—and so perform the function that they are supposed to perform. I share with Polanyi, Kitcher, and Goldman the belief that such judgments are not only essential but justified (and as far as I can tell, none of these authors would deny that they are imperfect in just the ways that I am about to describe).

I will consider four kinds of problems with authority judgments, what I call the *problem of determination*, the problem of the *poverty of evidence*, the *collapse of the orders*, and *network problems*.

3.1 Determination

The problem of determination is related to the generality objection to reliabilism. It poses the question: when I attempt to calibrate an authority's reliability, precisely what intellectual capacity am I measuring? For example, if I am assessing a scientist's experimental reliability—their ability to avoid mistaking experimental artifacts for real data—with respect to what class of experiments is that reliability being assessed? This is a question of practical relevance, I take it, because reliability will vary with the class: every kind of experiment brings its own complex set of challenges, and no one is equally competent at meeting all such challenges.

Assume that the determination problem can be solved in principle—that it is possible for a sufficiently accomplished calibrator to specify precisely the capacity whose reliability they are estimating. (This is perhaps doubtful, but let it pass for the sake of the argument.) Scientists nevertheless decline to solve it in practice. It is simply not worth their time and effort to pin down to this degree the significance of the authority-based weights that they calculate. Authority-based weight in real science is a nebulous thing, then: it is not an estimate of some well-defined quantity, but rather a placeholder whose worth lies in its being close enough to whatever such an estimate or family of estimates would be, were the issue of determination to be entirely resolved.

Let me emphasize that I take this imprecision or vagueness in the represen-

tation of authority-based weight to be the outcome of an entirely reasonable epistemic tradeoff. When it comes to authority in science, I am an observer, not a reformer. But it is a tradeoff all the same, and any scientist will have to recognize that as a result of accepting the deal, their assessments of authority have as their very subject matter something not entirely determinate.

3.2 Poverty of Evidence

The problem of the poverty of evidence arises in two ways: first, some scientists do not have enough of a track record for their reliability to be accurately assessed, and second, even where they have a track record (and they eventually, of course, accumulate one), it may be too much trouble, in many cases, to uncover the details. It is the latter difficulty that interests me here. Some scientist has published many experimental results, or many theoretical innovations, whose success could, with enough work, be quantified. But that scientist's total contribution to the pool of authority is not so great, so the return from determining the details of their career is also not so great. Even if there is someone out there who is familiar with these details—someone who would be an ideal calibrator—it may be too much work to find this person. In such cases, the scientist's reliability will be inferred from a proxy: the success of one or two papers, the prestige of the scientist's typical publication venue, or the prestige of their home institution.

A related problem: when indirectly calibrating—when using some other scientist's estimates of reliability to set your own—you ought to take into account not only their reliability as a judge of other scientists, but the extent of their knowledge of the particular case in question. In many cases it is, however, though possible, simply too much trouble to inquire into the judge's evidence base. In fact, one of the attractions of indirect calibration is the offloading of this sort of work onto other people. (This complication is a close neighbor of the network problems considered below.)

Two remarks. First, observe that, as so often in science, the poverty of

evidence is a matter of choice: given resource constraints, a cost-benefit analysis advises against collecting even what evidence already exists. Second, I do not want to push the case for evidential poverty too far. Frequently, there are excellent sources of information easily available that can be used to form a well-grounded estimate of a scientist's reliability. In matters where lives hang in the balance—tenure decisions, for example—such information will (or should) be sought out. But for more everyday decisions, it is just as frequently at the very least quite unclear how much, and what kind of, evidence underwrites an assessment of authority for a non-famous scientist.

3.3 Collapse of the Orders

The two remaining problems arise from difficulties in the accurate assessment of reliability due to the complexities involved in managing the process of second-order calibration, that is, the process of attempting to ascertain the likely accuracy of other scientists assessments of authority in order to arrive at a reliable indirect calibration.

The first of these two problems is the collapse of the orders. When taking into account other scientists estimates of reliability, you should weight them according to your estimates of those scientists' own reliability. Strictly speaking, these weights are not your estimates of the scientists' reliability as scientists—as experimenters or theorists—but are rather your estimates of their reliability as estimators. The weight you attribute to their opinions should, in other words, be proportioned to their facility in the second-order skill of arriving at accurate opinions of other scientists' authority.

To a certain extent, scientists (and other academics) do have such second-order opinions, distinguishing between a researcher's first-rate ability as an experimenter and their rather untrustworthy letters of recommendation. But more frequently, scientific skill is used as a proxy for the second-order skill. Because someone is a highly accomplished scientist, you accord their judgments of authority considerable weight when constructing your own.

Thus judgments of first-order and higher-order reliability are for heuristic purposes partially collapsed into one another.

This is an entirely justifiable practice. Scientific reliability is a decent enough proxy for the ability to assess scientific reliability, and information about scientific reliability is on the whole easier to come by—and so cheaper to obtain—than information about the second-order skill. Further, the use of the lower-order skill as a proxy for the higher-order skill provides a practical way out of a recursion problem: to indirectly calibrate estimators of scientific reliability, I would otherwise need to estimate scientists' ability to estimate reliably other scientists' ability to estimate first-order scientific reliability, and so on. So I do not question the rationality of the collapse of the orders. But it undeniably adulterates—to a degree—assessments of authority.

3.4 Network Problems

A network problem arises whenever there are unrecognized correlations among different points in the informational network. (See also Goldman (1999) on "failures of independence.") Three examples of the genre:

Double Counting I see that scientists *A* and *B* have endorsed a certain experimental result. Both are highly trustworthy, so I assign very high authority-based weight to the result, higher than if either scientist alone had endorsed it. Unknown to me, however, *B* endorses it only because *A* endorses it. I am in effect counting *A*'s endorsement twice.

Backscratching Scientist *A* issues a highly favorable endorsement of *B*'s work. *B* increases his estimation of *A* as a result, from its previous fairly high value to a very high value. (Assume that there is nothing inherently wrong with this.) Unknown to *B*, however, *A* was motivated to issue the endorsement of *B* only because she discovered that *B* had a fairly high opinion of her own work. *B*'s upward revision of that opinion is, then, in effect based on no

additional evidence. (I am using Kitcher (1993)'s notion of backscratching here, on which it is an entirely unintentional, purely epistemic phenomenon to be distinguished from logrolling, to be discussed next.)[3]

Logrolling Scientist *A* endorses an experimental result of *B*'s. As a consequence, I attribute some additional authority-based weight to *B*'s work. Unknown to me, however, *A* is angling for a job in *B*'s department, and has issued the endorsement without even reading *B*'s paper.

Two related remarks. First, network problems are, like the other problems here, solvable in principle. If I care badly enough, I can gather information about the relevant social background in sufficient detail to uncover the connections that create the kinds of misallocation of authority discussed above. But so often, although the evidence is out there, it is not worth the effort to collect it, because the increase in accuracy of my assessments of scientific reliability, thus of authority-based weight, will not repay the resources expended. Scientists face these problems as much by choice as by necessity, then—well-motivated choice.

Second, a range of interesting work on knowledge networks shows that network problems need not result in pathological epistemological distortions (Lehrer and Wagner 1981; Zollman 2007, forthcoming; see also note 3). But they are distortions all the same: they result in estimates of reliability, and thus assessments of authority, that are worse than they might otherwise be.

3.5 Dealing with the Provisional

In the light of all four problems surveyed here, and perhaps others, scientists should regard their estimates of authority with a certain degree of

3. Kitcher argues that backscratching will typically not result in an out-of-control feedback loop that drives *A*'s and *B*'s confidence in one another to the maximum value; it is, nevertheless, an epistemic distortion.

epistemic reserve. It is not merely that they should—as with any scientific hypothesis—suspend final judgment until more evidence comes in. With regard to authority assessment, their epistemic situation is precarious in two more unsettling ways than this.

First, because of the costs of information-gathering, representation, and computation, in judging authority scientists typically decide against taking into account even evidence that already exists. Thus their assessments of authority in a certain sense violate the principle of total evidence—though the violation is a rational one.

Second, as the determination problem shows, they have also often enough decided against thinking too hard about what hypotheses all of this evidence is evidence for. Thus they work with a space of hypotheses that is not entirely well defined.

In most of what follows, I will focus on the first reason for holding authority assessments at arm's length, that is, their disregard for the total evidence principle. The problem of the hypothesis space will, however, be briefly considered in section 4.6.

Is it correct to say that scientific authority assessments fail to respect the principle of total evidence? It is standard, of course, to formulate the principle in such a way that its violation cannot be rational, by interpreting the *available* in "Use all available evidence" very stringently, to apply only to evidence that is, as it were, staring you in the face *right now*. (Even then, it is not clear that adhering to the principle is compulsory once computation costs are taken into account.) Let me characterize a much more liberal—but also, I think, more idiomatic—notion of *available evidence*: evidence is available if it is there in the scientific journals, in scientists' curricula vitae, in letters of recommendation, in lunchroom gossip (if you have access to the lunchroom), or has otherwise been made public to some degree within the relevant scientific circles. In this sense, the track records of scientists, both in their roles as experimenters and theorists, and in their roles as first- and

higher-order calibrators, are largely available to you even if you choose not to take them into account. Likewise, information about political alliances and so on is, in this same sense, for the most part (though not entirely) available.

In science's assessment of authority, I am claiming, much available evidence goes unused—a great contrast to science's assessment of hypotheses, in which all or almost all available evidence is typically taken into account. It is for this reason, I propose, that scientists should take a somewhat different, and more wary, epistemic attitude to assessments of authority-based weight than they do to what might be called the "scientific facts," or what I will later call the "evidence structure" (section 6). That is, a kind of mental or institutional question mark—though a subtle one—should hang above assessments of authority in science. Not the kind of question mark that should prevent scientists from putting those assessments to use, but one that . . . well, that is the next question. What ought to be the practical consequences, if any, of the recognition that assessments of authority are not formed in accordance with the principle of total evidence?

I see four possible answers to this question:

1. In assessing authority-based weight, make use of all available evidence.

2. Expunge authority assessments from science.

3. Ignore the question mark; go ahead and make use of the best assessments of authority you have as you would make use of any other beliefs or probabilities in science.

4. Something between (2) and (3): Do not neglect judgments of authority, but treat them differently from other judgments made more in accordance with the requirement of total evidence.

The first two options—the purist options—cannot be taken seriously. Answer (1) is off the table because it is, for reasons given above, irrational to make use of all the available evidence: the costs of collection and computation

exceed considerably the net improvement in assessments of reliability. (It is perhaps not so much that the net improvement is small, but that the costs of information gathering and processing are extraordinarily high.) Answer (2) is not just off the table, it is not even in the lab: modern science without deference to authority would be impossible.

Answer (3) is worth a closer look. Our assessments of authority may be flawed in various ways, but they are the best assessments that we have—and in the short term, the best assessments that we are likely to have, given our cost-benefit-based decision to decline to improve them. We need to make various authority-based decisions in the same short term, so what objection can there be to putting them to use?

The one major argument against this strategy stems from the following observation: Available evidence that at one point seems too much trouble to take into account may, in the light of shifting priorities or new information, become worth a closer look. We would like to be able to go back and adjust our assessment of the relevant authority in the light of the unused evidence, and we would then like to update our total epistemic state to reflect this adjustment. I will call such a process *epistemic backtracking*.

To allow backtracking is to treat authority assessments, and thus to treat epistemic probabilities based on authority assessments, as provisional. On the assumption that not every factor that modulates belief change in science is provisional in exactly the same way, then, it is to treat authority assessments differently from other epistemic factors in science. That requires the rejection of option (3) and the embrace of (4), so setting the stage for the rest of this paper, in which several semiformal approaches to backtracking are considered.

⋆ ⋆ ⋆

Before I continue, let me clarify the nature and force of the argument for allowing a scientist to backtrack.

First, a backtracking reasoner need not be timid in the use of their au-

thority assessments. Though a backtracker reserves the right to revise their assessments in the light of changing contexts and priorities, they believe that their current assessments make optimal use of the available evidence; they should not, then, be shy in applying their authority judgments where necessary—and in science, such applications are frequently necessary.

Second, backtracking need not be confined to authority judgments. Whenever evidence is available but unused—for whatever reason, but presumably most often because the costs involved outweigh the likely epistemic benefits—the occasion may arise to go back and take advantage of some of the unused available evidence, or in other words, to backtrack. For example, a linguist advancing a hypothesis about universal grammar might not consult all published facts about the world's many languages. But as the hypothesis is challenged from various quarters, it might become clear that some particular language or language family will advance the issue, at which point this available but previously ignored piece of evidence can be incorporated into the calculation of the hypothesis's empirical standing. What I have to offer in this paper can be applied to such cases; it is, then, an approach to problems arising from the practical bounds on scientific rationality in general, or at least an approach to those problems arising from bounds that limit our ability to honor the requirement of total evidence. I will, however, confine my attention to assessments of authority, where I believe that the total evidence requirement is evaded far more frequently and far more profoundly than elsewhere in science; the generalization of the approach is left to the reader.

Third, and also for the purpose of delimiting my inquiry, let me say something more about the meanings of my terms of art. As I define it, *backtracking* with respect to your opinion about a hypothesis is possible just in case you are taking account of available evidence that you previously ignored. Your attitude to a hypothesis is *provisional* just in case some available evidence was ignored, thus just in case backtracking with respect to your opinion about the hypothesis is possible. The definitions of both terms

are rooted, then, in the definition of *available evidence*. That notion might be narrowed or broadened to yield different notions of backtracking and provisional opinion. For example, on one narrower definition, evidence is "available" only if it is at your fingertips; it may reasonably be ignored (in normal circumstances) then, only because of computational costs. On a wider definition, you might count evidence as "available" if the experiments that would yield it are in some sense extremely easy to perform; "backtracking," then, would consist of going back to make the observations in question. Any of these different ways of fleshing out the notion of available evidence, and thus the notions of backtracking and provisional opinion, might be useful in the study of bounded rationality, by way of the semiformal implementations of backtracking that I will discuss in what follows. But I will not depart from the definitions induced by the (admittedly imprecise) sense of availability introduced above, on which evidence is available just in case it has been made public to a sufficient degree in a scientific context.

Fourth, a few more words about the rationale for violating the principle of total evidence, that is, for declining to make use of all available evidence relevant to a given hypothesis. The considerations that go into such a decision might be divided into several classes:

1. Most obviously, the cost of obtaining the information and the size of the difference it is likely to make to your epistemic state.

2. The intrinsic importance of having a correct scientific opinion about the hypothesis. This importance might change either for practical reasons, as human needs change, or for epistemic reasons, as the needs of other parts of science change.

3. The point at which the correctness of scientific opinion about the hypothesis becomes important. In some cases, that opinion will matter in the short term, as when the hypothesis concerns matters of immediate practical significance. In other cases, the opinion will matter for the

most part only in the longer term. How is this distinction important? In the longer term cases, there will typically be less reason to take into account all available evidence, since the differences in opinion with and without the evidence can be expected to be "washed out" by further evidence to arrive in the future. Some available evidence will be justly ignored, then, because it makes no long-term difference.

4. Strategic considerations. It might, for example, be preferable for scientists not to take into account too much information about their colleagues' beliefs about a problem, if a few authoritative pronouncements would stifle much-needed diversity in the range of approaches to the problem. In such a case, efforts to ensure short-term correctness impede the prospects of long-term correctness (a possibility that will matter only, of course, if short-term correctness is relatively unimportant). The question how to tune attention to authority in the short term so as to find a level of diversity that maximizes correctness in the long term has been explored with considerable insight by Zollman (2007, forthcoming).

There is, clearly, much more to say about these decisions; it will not, however, be my topic in the present paper. I will simply assume that it is reasonable, in some circumstances, to ignore available evidence, and equally reasonable, under other later circumstances, to go back and take into account what was previously ignored.

4. Qualifying Authority

I have argued that the adoption of some sort of backtracking technique is the most reasonable response to the frailty of authority assessments, and in particular to their failing to take into account all the available evidence. A

backtracker is, in effect, reserving the right to go back and look more closely at some of that evidence.

How ought backtracking to be implemented, then? The question has two parts. First, how should the special, provisional status of authority assessments be represented? How to single out authority-based weights so that any epistemic probability calculated on the basis of such weights knows its authority-based origins and comes in, when needed, for a retrofit? Second, how should the process of backtracking itself be implemented? When an authority-based epistemic probability is altered, how to ensure that the effects of the alteration percolate through your total epistemic outlook? These questions need to be tackled in tandem; let me consider some answers, drawing inspiration from the extensive literature on uncertainty in formal epistemology, statistics, and economics.

I consider four approaches. The first (section 4.1) involves no backtracking at all; a tentative attitude to authority-based weight is represented in a quite different way. The second (section 4.2) and to some extent the third (section 4.3) treat old available evidence newly taken into account in the same way as new evidence; they make no deep distinction, then, between backtracking and any other way of learning from the evidence. The fourth approach takes backtracking seriously as a distinctive epistemic process.

4.1 Qualification without Backtracking

I will begin with a very simple form of qualification that seeks to do away with the need for backtracking altogether.

The idea is to apply a discount to authority-based weight so that it has less impact than it otherwise would. Let me motivate this idea by considering a simple scenario. Take two epistemic probabilities attached to two rival hypotheses. Suppose that each of the probabilities is about half attributable to directly observed empirical evidence and half is attributable to authority-based weight. (Never mind how such an attribution is possible.) Then in

making judgments based on the probabilities attached to the hypotheses—in choosing which one is worthier of further investigation, say—you might think that, in the light of the somewhat shaky status of assessments of authority, we should pay more attention to the half based on direct observation than to the authority-based half.

There are two ways to apply this discount: before and after. "After": you keep track of the proportion of an epistemic probability that is attributable to authority-based weight and discount this authority-based portion at the time of decision making. "Before": you apply the discount at the time that authority-based weight is first taken into account, that is, when the epistemic probability is first adjusted in the light of the assessment of authority.

In this section I will be interested in the "before" strategy,[4] a strategy that in effect gives you two different methods for taking new information into account. One of these methods—the method you use when learning from authority—is simply, as it were, more conservative: it gives the same amount of weight less impact on your pre-existing epistemic probabilities than the other method would give it. As Carnap (1950) might have formalized it, you use a higher lambda value when taking into account opinion than when taking into account fact.[5]

The "before" discounting technique has the considerable advantage of requiring no bookkeeping—you do not need to keep track of authority-based weight once it has entered into your epistemic calculations. It has, I think, two disadvantages.

The first, more philosophical disadvantage is that it verges on incoherence. It is surely constitutive of epistemic weight that all pieces of information with

4. The system for keeping track of provisionality described in section 4.4 might be used to construct an "after" approach.

5. Carnap's inductive logic has a single parameter λ that determines the speed at which you abandon prior beliefs in the light of new information. The higher the value of λ, the more conservative your inductive strategy. At the extremes, when λ is zero, you give no weight at all to your prior beliefs, whereas in the limit as λ goes to infinity, you give no weight at all to the evidence.

the same weight in respect to some hypothesis have the same impact on that hypothesis's epistemic probability. If I trust one source of information less than another, that fiduciary differential ought to be reflected in my giving relatively lower weight to the former than to the latter, not in my treating weight from the two sources differently.

Perhaps, then, the suggestion ought to be to give authority-based information less weight than it would otherwise have? How to conceive of such an operation? Do I calculate the weight that the information ought to have, then discount it? That is surely irrational: I am attributing to evidence some weight other than the weight it ought to have. Then it must be that the discount is already built into the weight that ought rationally to be attributed to authority. That is not a bold strategy, however, but a platitude: authority-based information ought to be attributed the weight that it rationally deserves. As such, it is no help at all with the problem posed in this paper, that of how to deal with the fact that, even once the proper weight of authority-based information has been determined, it ought to be regarded, in the light of (among other things) its not taking all available evidence into account, as provisional. In effect, the discounting solution's problem is that it conflates a property of weight—its provisionality—with weight itself. The other approaches to qualifying authority considered in this paper attempt, by contrast, a distinct representation of provisionality.

The second and more practical disadvantage of the discount schema is that it does not treat the provisional status of weight intelligently. When some hypothesis or experiment turns out to be far more important than we supposed, we want to reconsider our authority-based assessment of the matter, taking into account more of the available evidence than it previously seemed worthwhile to consider. The discounting scheme does not allow this. In some cases, perhaps, the implicit tradeoff is worthwhile: we behave in an epistemically less sophisticated way, but at an enormous saving in time and bookkeeping effort. But in other cases the tradeoff is surely a bad deal; for

those cases, we need something more refined.

Suppose, for example, that two experiments yield evidence bearing on a hypothesis, one in favor and one against. At the time, the hypothesis seems unremarkable, and so you conduct only a cursory assessment of the authority of the experimenters, rating them roughly on a par. The result is an epistemic stalemate. Later, the hypothesis turns out to be very important. It makes sense, in the light of the provisionality of your initial judgments, to go back to look at the experimenters' credentials more closely. But the discounting scheme has already taken the provisionality into account by reducing the weight of both experiments equally, giving you no means with which to distinguish between the experiments, even after taking a closer look at the available evidence.

4.2 Bayesian Tracking

Bayesian epistemology provides a straightforward way to track the uncertainty or provisionality of an epistemic probability. Although I call this method "Bayesian tracking," it is not proprietary to Bayesianism—it is available to any epistemology that attaches epistemic probabilities to propositions. Nor is it the only way that Bayesians might deal with the problem of authority, as you will see below. But it is the natural Bayesian method for representing the provisional nature of authority-based assessments.

Let me work with a classic example. Consider three witnesses to a coin toss. Each assigns an epistemic probability of one-half (that is, in the Bayesian system, a subjective probability of one-half) to the event of the coin's landing heads. But they differ in the degree to which they are confident in this epistemic probability, that is, they differ in the degree to which they regard the probability they assign to heads as settled.

The first witness knows that the coin is fair, and thus that the physical probability of heads is one-half. Short of a crystal ball or some such thing, no evidence would move her to alter her epistemic probability for heads of

one-half.

The second witness does not know that the coin is fair. She does know that it was selected at random from an urn containing a range of coins, some fair, some biased toward heads, and some biased toward tails. She knows the exact distribution of such coins, so she knows the physical probability that the chosen coin is fair, that it is biased to give heads two-thirds of the time, and so on. In particular, she knows that the biases are balanced; this is why her epistemic probability for heads is one-half.

The third witness, like the second witness, does not know that the coin is fair but knows that the coin was selected from an urn containing a range of coins, some biased and some not. Unlike the second witness, she does not know the distribution of coins in the urn. For all she knows, most of the coins are biased toward heads. Or perhaps they are mostly biased toward tails. Then again, perhaps, they are almost all fair. She assigns epistemic probabilities to these different possibilities that balance out, yielding an overall epistemic probability for heads of one-half.

The first witness's epistemic probability of one-half is (prior to the toss) impervious to almost all evidence. The second witness's epistemic probability is subject to change if a certain kind of information is received, namely, information about the bias of the coin used for the toss. (I say "subject to change" because in the special case where the coin is fair, there will be no actual change.) The third witness's epistemic probability is subject to change on receipt of a wider range of information—either information about the bias of the coin or about the composition of the urn. It is in this sense that the third witness's epistemic probability for heads is less settled than the second witness's, and the second witness's is less settled than the first witness's.

This unsettledness does not show itself in the value of the epistemic probability itself, which is the same for each of the three witnesses. It shows itself rather in certain related epistemic probabilities. Where the first witness has an epistemic probability of one (or close enough) that the physical probability

of obtaining heads on the selected coin is one-half, the other two witnesses do not. And where the second witness has an epistemic probability of one that the urn has a certain composition, and thus an epistemic probability of one for a certain physical probability distribution over the bias of the selected coin (determined by the composition of the urn), the third witness does not. Thus the first witness's epistemic probabilities are concentrated in a very small subset of the space of possibilities, the second witness's epistemic probabilities are concentrated in a somewhat larger superset of this subset, and the third witness's on a larger superset still. In each case, half of the epistemic probability distribution sits, as it were, on the event of *heads*, so the witnesses assign the same probability to heads, but their differing levels of certainty, or of settledness, are represented by the spread of the probability distribution over the relevant area: the wider the probabilities are spread, the less the settledness and so the more the uncertainty.

The observation generalizes: the less settled your epistemic probability for a certain possibility—the wider the range of evidence (available or otherwise) that might cause a change in that probability—the wider the spread of the epistemic probability distribution in the relevant dimensions.

This provides, then, a more or less automatic solution to the problem of representing the provisionality of authority assessments, and of reevaluating those assessments in the light of new evidence. It works as follows. Suppose that you are assessing the authority of some experimenter (with respect to some particular class of experiments). Suppose further that your assessment is based, for the most part, on what Kitcher calls an indirect calibration: you have assembled the views of local experts and constructed an aggregate assessment of the experimenter, weighting the experts' individual assessments according to your estimate of their reliability. For the reasons given in the previous section, you will be unsure, typically, about a range of things, including:

1. The experts' assessments (that is, the precise degree of authority they

attribute to the experimenter),

2. The reliability of the experts' assessments,

3. The independence of the experts' assessments (to what degree are they relying on one another?).

This uncertainty, and thus the provisionality of your own indirect assessment, will be reflected in your epistemic probability distribution's being spread over a range of possibilities for each of the factors above. For example, it will spread over a number of different possible values for each expert's reliability, reflecting your uncertainty as to exactly how reliable they are.

How does backtracking work? If it becomes worthwhile to go back to examine more of the available evidence about (say) the experts' reliability, you will not have to do anything methodologically novel; you simply take the evidence into account in the usual way, and your epistemic probability distribution over the possibilities will narrow as a result. (I ignore the problem of old evidence; see section 4.5.) Compare the case of the coin toss: if the third witness learns the distribution of the coins in the urn, her distribution will contract to the size of the second witness's distribution; if she then learns the bias of the selected coin, her distribution will contract further to the size of the first witness's distribution. As the comparison shows, Bayesian tracking treats unused available evidence like new evidence; strictly speaking, then, it involves no backtracking at all: you are not "unconditionalizing" and then conditionalizing again with new probabilities; there is no epistemic "rewinding."

Such a method for tracking and revising authority assessments is very elegant in theory. But it requires a great deal of bookkeeping. In the case where the unsettledness of an assessment is due to a principled disregard of available evidence, in particular, the effort required to maintain well-defined probability distributions over all the relevant possibilities is probably not a great deal less than the effort that would be required to collect and take

into the account the available evidence itself. After all, you need to represent separately every possible way in which the opinions of the experts might be interconnected, including all permutations of backscratching, logrolling and so on (see network problems above), every precisification of the capacity that is to be assessed (see the determination problem above), including all the higher-order capacities (see the collapse of the orders above). It was precisely to avoid this kind of expenditure of effort that you allowed your original assessments of authority to remain provisional.

Perhaps a bulging toolbox of principles of indifference could help you to construct the necessary array of epistemic probability distributions? Such principles churn out a probability distribution once the set of possibilities is specified, but the the greater part of the bookkeeping task that confronts the Bayesian tracker is representing these possibilities distinctly to begin with.

In short, Bayesian tracking is a lovely idea, but the cost of maintaining such an elaborate representation of your epistemic state is surely too high to be worthwhile—unless, perhaps, there is absolutely no other way to deal intelligently with the provisionality of authority assessments. But there is.

4.3 Second-Order Probability

Can the uncertainty or provisionality of your epistemic probability for a hypothesis be represented by a second-order probability qualifying that epistemic probability?

Some background. A second-order probability is a probability that some first-order probability is in some sense "correct." Probabilities at the two levels may or may not be of the same type. In the example of the coin toss in the previous section, for example, there are both physical probabilities for physical probabilities (the physical probability that a fair coin is chosen from the urn is also a physical probability that the physical probability of obtaining heads on the chosen coin is one-half) and epistemic probabilities for physical probabilities (witnesses 2 and 3 have epistemic probability distributions over

the different possible physical probabilities for heads; witness 3, note, also has an epistemic probability distribution over the second-order physical probability distributions—a third order probability distribution). Epistemic probability distributions over physical probabilities are the basic stuff of modern Bayesian confirmation theory; the viability of the Bayesian and other systems should convince you that there are no special formal problems in setting up higher-order probability distributions.[6]

How, then, to use second-order probabilities to qualify assessments of authority-based weight? The problem of authority assessments has been cast, in this paper, as arising from a (rational) violation of a kind of principle of total evidence: authority assessments should be qualified because they typically do not take all available evidence into account. The assumption underlying the problem, when it is set up in this way, is as follows. Were you to be presented with all the available evidence relevant to a particular authority assessment, along with the time and computational means needed to take it into account (but let me not make an issue of this computational factor in what follows), you would have no reason to treat the assessment differently from your other epistemic probabilities. Thus, for any given scientific hypothesis whose epistemic probability depends on some particular set of authority assessments, were you be presented with all available evidence relevant to making those assessments, you could adopt an epistemic probability for the hypothesis that needed no special qualification. Your qualification of an epistemic probability is entirely derived, in short, from your (self-imposed) uncertainty about the available evidence. The suggestion I examine in this

6. There need not be any interaction between the different orders of probability at all; typically, however, there is a coordination principle linking the orders. When first-order probabilities are physical and second-order probabilities are subjective, for example, something like Lewis's "Principal Principle" (also sometimes called Miller's Principle) does the job, by stating roughly that your subjective probability for an event, conditional on the physical probability of that event's taking the value x, should also be x. (Lewis (1980) laid out some of the important ways in which this principle must be qualified; work on the content of the correct qualifications continues (Strevens 1995).)

section is that your qualification should be represented by a second-order probability distribution representing that uncertainty.

The first-order probabilities in this scheme will represent the epistemic probabilities you would assign to a hypothesis were you to know that the available evidence was such and such, for each possible set of available evidence. Such a probability is the correct probability for the hypothesis just in case the available evidence is as supposed. The second-order probabilities are a distribution over the different possible sets of available evidence.

More formally, for any possible complete set of available evidence a, there will be a first-order probability distribution $P_a(\cdot)$ over the hypotheses and other propositions that gives your epistemic probability of a proposition on the assumption that the available evidence is a. Because it is not worth your while to take into account all available evidence, you will typically not know the composition of the actual set of available evidence; rather, you will have a probability distribution $C(\cdot)$ over the different possible candidates. These two probability distributions are linked by a coordination principle, for which the following is the simplest possible form:

$$C(h \mid k_a) = P_a(h)$$

where k_a is the proposition that the complete set of available evidence is a.[7] More generally, your epistemic probability $C(h)$ for a hypothesis h will be a weighted average of the first-order probabilities $P_a(h)$. The mathematical structure of the system, then, is analogous to the standard Bayesian system in which there is a single subjective probability distribution over different possible physical probability distributions, with $C(\cdot)$ playing the role of the subjective probability distribution and the various $P_a(\cdot)$s playing the role of the physical probability distributions. In both systems it is the second-order distribution $C(\cdot)$, informed by the appropriate coordination principle, that is

7. Questions as to how to incorporate background knowledge and so on into the principle will depend on the variety of epistemic probability involved.

your workhorse—it is what you use to decide what experiments to perform, what papers to publish, what public policy to advocate, and so on.[8]

The hierarchical probability structure represents uncertainty about a hypothesis by representing uncertainty about the contents of the set of available evidence. In other words, uncertainty is represented by the spread of the second-order probability function $C(\cdot)$ over the different candidate sets of available evidence—just as you would like. Epistemic backtracking is straightforward: when you decide to investigate the basis of some authority assessment further, you gather information about the evidence basis for that assessment, reducing your uncertainty about the available evidence. Your second-order probability distribution contracts, and your overall probability for the hypothesis changes accordingly, reflecting your new judgment.

Some readers will have noted that the hierarchical system is very similar, in the way it keeps track of provisionality, to the Bayesian tracking described above. And indeed, it shares Bayesian tracking's virtues and its vices. On the one hand, its implementation of backtracking is elegant and comprehensive (without requiring any actual "rewinding", although it does make a formal distinction between taking into account previously available evidence and uncovering new evidence). On the other hand, it makes intolerable cognitive demands of the backtracker. To put the system to work, you must represent every single way that the available evidence might come out, along with a probability for the hypothesis of interest calculated on the assumption that

8. The hierarchical probabilistic approach of which the above suggestion is an instance has been applied in a number of ways to a number of problems in probabilistic epistemology. In the philosophy-friendly literature, Good's (1980) contributions are especially notable; see also Skyrms (1980). More recently, Williamson (2000) has advocated a probabilistic representation of an epistemology with an externalist conception of evidence—a conception on which you are not always sure what your evidence is. Such a representation puts a set of first-order probability distributions over propositions that gives, for any particular posit about your total evidence, the probability of the proposition on that evidence, and a second-order probability distribution over the different possible sets of total evidence. It is therefore structurally very similar to the proposal currently under examination, though employed in the service of somewhat different philosophical ends.

it does come out that way. This responsibility is not quite as taxing as in Bayesian tracking, where you must represent every possible state of the world, not merely every possible set of available evidence, but it is enough to put backtracking using second-order probabilities over the available evidence, attractive though it may be, entirely beyond scientists' grasp.

To put it another way, although this highly idealized picture of the way in which we deal with uncertainty may have its epistemological uses, it is not of much use for inquiring into the advantages and disadvantages of different backtracking schemes, since the very factors that make backtracking necessary have been more or less idealized away.

4.4 Second-Order Nonprobability

Second-order quantities attached to probabilities need not themselves be probabilities. Suppose, for example, that you qualify any particular first-order epistemic probability for a hypothesis with a further quantity that represents the proportion of available evidence taken into account when calculating that probability. This is not a part of a second-order epistemic probability distribution, because it does not, when combined with some complementary quantity attached to another first-order epistemic probability, sum to one.[9] It does, however, quantify a property of the probability, so it is right and proper to think of it as a second-order entity in your epistemic hierarchy.

Keynes (1921, chap. 6) discusses the kind of second-order quantity I have in mind, or a rather a more general counterpart: his quantity measures the amount of evidence taken into account when assessing a first-order probability relative to, not just the available evidence, but to the total *possible*

9. The only candidate complementary quantity is the proportion of available evidence *not* taken into account when calculating the original first-order probability, but this is not attached to any other first-order probability. That said, because it is a proportion, the quantity does satisfy the axioms of probability—it is formally a probability distribution, but not a distribution over your first-order epistemic probabilities.

evidence, including unknown facts and facts that have yet to come to be. Borrowing a term from my discussion of Bayesian tracking, call this quantity a first-order probability's *settledness*.[10] Keynes asks whether settledness has any direct decision-theoretic significance: given the choice between playing two distinct gambling games with the same expected value, is it better, when all other relevant factors are equal, to play the game whose expected value is based on epistemic probabilities of greater settledness? That is, should you prefer to take chances when your calculation of the odds is based on greater rather than lesser amounts of evidence? Keynes is unable to decide the question. Subsequent writers have tended to think that settledness should not make a difference—it plays no role in standard decision theory—but Ellsberg (1961) famously demonstrated that ordinary humans do in fact show the preference that Keynes described. As economists sometimes say, the folk prefer "risk" to "uncertainty" (aka "unsettledness"). Some attempts to model the decision-making principles underlying the "Ellsberg paradox" posit a second-order representation of settledness of the sort suggested by Keynes.[11] It is such a representation that interests me here.

Suppose, then, that attached to every epistemic probability is a quantity measuring something like the amount of available evidence taken into account in calculating the probability. Since the quantity cares about available evidence only, I will not say that it measures settledness or uncertainty but rather *provisionality* (making quantitative the term of art defined in section 3.5). The more available evidence is taken into account in determining the probability of some hypothesis, then, the less provisional the probability, or in other words, the more available evidence is ignored in determining the probability, the greater its provisionality (ceteris paribus).

I will not decide among many possible approaches to characterizing and

10. Keynes calls his quantity *weight*, a term that has since acquired a rather different meaning in the epistemological literature, roughly synonymous with "degree of evidential relevance." It is in the latter sense that I use the term "authority-based weight."

11. For a survey of these and other related techniques, see De Cooman and Walley (2002).

calculating provisionality. For example, I will not say whether provisionality measures the absolute quantity of available evidence ignored, the absolute quantity of available evidence taken into account, the proportion of available evidence taken into account, or something else. I will not worry about uncertainty in our estimates of what is ignored, and thus our (third order!) beliefs about the provisionality of our estimates of provisionality. I will not choose between different methods for updating the provisionality of a hypothesis when new evidence, also provisional, comes to light (though see note 13). Nor will I say anything about the relation between the provisionality of a compound proposition and the provisionality of its parts.[12]

Assuming that the system of provisionalities is set up properly, you now have a running tally of how provisional any particular hypothesis is, without having to do a great deal of bookkeeping. If a hypothesis becomes very significant but its provisionality is dangerously high, you can go back and collect more relevant information—that is, more information relevant to determining the authority assessments that went into your original epistemic probability for the hypothesis. Your provisionalities therefore give you a sound basis for deciding when to backtrack. But they are of less help in deciding how to backtrack.

To see this, take a closer look at the way in which provisionalities are updated. Suppose I receive a new piece of evidence e for a hypothesis h. Suppose further that the two epistemic probabilities that matter in taking this evidence into account are the (posterior) probability of e and the conditional probability of e given h (that is, the likelihood of h on e). Finally, suppose that both probabilities are in part authority based and so have a certain level

12. On this last question: is the provisionality of a logical formula always a simple function of the provisionality of its components? Can the provisionality of ab always be derived directly from the provisionality of a and b? If there can be evidence for or against ab that is not evidence for or against a or b in isolation, then the answer is surely negative—though it may be that in the vast majority of cases no such evidence is actually available, so that the list of exceptions to a simple and systematic rule would be short and manageable.

of provisionality—the probability of *e* because the experiment is a tricky one and so the ability of the experimenter is an issue, and the likelihood because it depends on auxiliary hypotheses that are themselves to some degree controversial.

I adjust my probability for *h* in the light of the new evidence. As a consequence, this probability presumably inherits some of the provisionality in the two probabilities that determine the adjustment, that is, some of the provisionality in the probability of *e* and the likelihood of *h* on *e*. I am assuming as I noted above that there is some rule telling me how to update the provisionality as I update the probability; never mind how it works.[13]

Now suppose that *h* later turns out to be more important than I realized when I made my cost-benefit decision as to how much time and money to spend on the authority assessments that went into determining its epistemic probability. In the light of *h*'s new status, its provisionality is unacceptably high. What then? I want to unravel some of the calculations that went into determining *h*'s probability, and in particular, I want to rethink some authority assessments. But which? I would like to (a) track down the prime sources of *h*'s provisionality, that is, the epistemic probabilities on which *h*'s probability depends that did the most to raise *h*'s probability's provisionality; (b) find

13. But of course this is a very interesting question! If provisionality is defined as the proportion of the relevant available evidence taken into account, then it seems clear how the provisionality must be updated: I add the available evidence pertaining to the newly relevant probabilities to the denominator, and the amount of that evidence that I have taken into account to the numerator. (It seems, then, that I must keep track not just of the the fraction itself, which I have been calling the provisionality, but the numerator and denominator of the fraction.) Then again, I might treat provisionalities like lower probabilities (or, depending on how you look at it, upper probabilities), taking the provisionality of *h* to be the maximum of (a) its previous provisionality, (b) the provisionality of the probability of *e*, and (c) the provisionality of the likelihood of *h* on *e*. There are numerous other possibilities.

A related question: to what extent should you keep track of the particulars of the evidence taken into account? In principle, you should want to keep track not only of the amount of available evidence taken into account when calculating the probability of *h*, but also the content of the evidence, so as to avoid double-counting if the same piece of evidence becomes relevant in two different ways. In practice, there will surely be ways to avoid any significant degree of double-counting that do not require so much clerical work.

enough evidence to lower their provisionality; and then (c) recalculate my probability for *h*. The system of provisionality does not help me with any of this: it keeps track of current provisionality, but it leaves no paper trail. Thus, it helps me to see that I need more information about *h*, but does not tell me which information.

The same observation applies to various other schemes for tracking uncertainty, such as the use of interval probabilities ("vague" probabilities) to represent uncertainty as to a probability's correct value, because they too have no memory, nothing functionally equivalent to a map of where you have been epistemically to help you find your way back so as to do everything—or at least something—all over again, in the light of your new priorities.

⋆ ⋆ ⋆

I criticized Bayesian tracking and certain schemes of second-order probability on the grounds that they force the scientist to maintain too fine-grained an epistemic state concerning too many possibilities. Now I am criticizing other second-order schemes on the grounds that they do not record enough epistemic information—that they are epistemically too coarse grained. Is there a happy medium? Or is there a stark choice between, on the one hand, epistemic backtracking at the expense of enormous cognitive effort, and on the other hand, practical but simplistic handling of the problems introduced by scientific authority?

From a purely psychological perspective, it is clear that there must be a middle way. I can keep a journal of my probabilistic calculations that lays out the reasoning that goes into calculating my epistemic probability for a hypothesis step by step. The journal will document the points at which authority assessments enter into the calculations. If some of these assessments later come to be seen as inadequately evidenced, I can go back, find more evidence, update the assessments, then recalculate my probability for the hypothesis using the same techniques and the updated values.

The problem is to find some description of this process that is both

sufficiently formal, or at least sufficiently systematic, that its rationality and reliability can be usefully discussed, and sufficiently cognitively feasible that it has a some hope of ethnological reality—that is, sufficiently feasible that it might capture present-day scientific practice, or provide the blueprint for some superior future scientific practice.

An obvious next step is to turn to anthropology, to see how scientists actually handle backtracking. But before I stoop to inspecting the empirical facts, I will go to the formal epistemological well one last time to see what the bucket brings up.

4.5 New Theories and Old Evidence

Something resembling epistemic backtracking has been proposed as a solution to Bayesian confirmation theory's problem of old evidence (or as it might perhaps be more perspicuously called, the problem of new theories). Let me take a closer look.

Suppose that a new theory is formulated to predict and explain phenomena already known to science; Glymour (1980)'s paradigmatic example is Einstein's general theory of relativity (GTR). When GTR was conceived, many of the facts about the orbits of the planets predicted by the theory were already known. In particular, the precession of the perihelion of Mercury's orbit was already known; at the time it had no convincing Newtonian explanation, and Einstein's explanation of the precession was considered a major evidential coup for his new theory.

The Bayesian approach to confirmation has trouble accounting for this fact.[14] An ideal Bayesian treatment of the historical episode would be as follows. Upon formulation, GTR is assigned an initial subjective probability. The existing evidence is then brought to bear on the new theory: GTR's initial

14. Technically, the problem is that the Bayesian system assumes that all possibilities to which probabilities may be assigned are known in advance; there can be no such thing as a new theory, then.

probability is adjusted according to whether it probabilifies this evidence to a greater or lesser degree than its rivals. Since Mercury's orbital peculiarities are predicted by GTR but not by other known theories of gravitation, they lend strong confirmation to GTR.

But in Bayesian confirmation theory, any attempt to conditionalize on a piece of evidence that is already known will have zero effect on your probability distribution: Bayesianism allows evidence to be brought to bear on a hypothesis only at the time that the evidence is uncovered—only at that moment that its subjective probability leaps to one. In the case of the precession of Mercury's orbit, this moment has already passed; the fact of the precession can have no effect on the probability of GTR. As a theoretical latecomer, GTR has missed its chance at confirmation. If only scientists had spent less time with their telescopes and more with their tensors . . . [15]

This is not a paper on old evidence, so I will not survey the various solutions to the problem. I am interested in only one of these: the suggestion that, to allow GTR to experience the full benefit of its gravitational savvy, we should in some way rewind our epistemic history to a point before the discovery of the precession of Mercury's orbit and conditionalize on the evidence all over again, as if for the very first time. I will examine a particular version of this approach that takes the idea of an "epistemic rewind" more seriously than most.[16]

We are working within the Bayesian framework, so before any evidence arrives, assign prior probabilities to the theories that you wish to test. These are what I will call your *initial priors*. Divide all probability among the

15. As I understand the problem of old evidence, then, it consists in the Bayesian's inability to bring old evidence to bear on a newly conceived theory by conditionalization—it is what Garber (1983) calls the historical problem of old evidence, not what he calls the ahistorical problem of producing a theory of timeless evidential support.

16. A version that Glymour considers (and rejects) looks explicitly to historical values for the prior probability of the evidence, while Howson's (1984) solution has an ahistorical air, and perhaps ought to be regarded as a solution to Garber's "ahistorical problem" (see note 15). What I offer here takes its cue from Skyrms's (1983) idea of "keeping a diary."

known theories. Note that there is no "catch-all" hypothesis to capture the possibility that some unknown theory is correct—the appearance of new theories will be handled dynamically. (I will have more to say about this feature below.) Remember the values of these initial prior probabilities; they will be consulted when backtracking occurs. Now open the empirical shutters and let the evidence shine in. As it arrives, make a note of it—you are keeping a record of all your evidence—and then conditionalize as usual using Bayes' rule.

Suppose that a new theoretical possibility comes to light, a GTR or such-like. Go back to your initial priors, that is, the priors you assigned in your pre-evidential state. Assign whatever prior you like to your new theory; reduce the priors of the preexisting theories accordingly. You now have a new set of initial priors. Using these new priors, conditionalize on all the evidence all over again. Repeat as necessary.

What are the informational and processing costs of this backtracking? There are several sets of facts that will be needed for your next recalculation, and so concerning which you should keep records:

1. Your most recent set of initial priors, that is, your prior distribution over all theories currently on the table. (No need, then, to keep track of assignments of initial priors that predate the appearance of your most recent "new" theories.)

2. All the evidence that you have received.

3. The likelihoods of the theories relative to each piece of this evidence.

4. The prior probability of each piece of evidence immediately before it is reported.

I assume that the evidential probabilities (4) are calculated when needed from the other available probabilities, using the theorem of total probability; no

separate record need be made of their values. (Having no catch-all hypothesis is important for the feasibility of this approach.)

The main informational load, then, consists in keeping a record of (a) all evidence received and (b) whatever means are necessary to calculate the likelihoods of the live hypotheses on that evidence. The means of calculation (b) will specify all relevant auxiliary hypotheses along with their role in deriving the probability of the evidence; it will not, however, include a record of prior probabilities for the auxiliaries. Thus, in the case where there are rival auxiliaries, what is recorded is the physical probability bestowed on the evidence by each combination of auxiliary and main hypothesis; because priors for the auxiliaries are not recorded, however, no value will be specified for the likelihood of the hypothesis simpliciter on the evidence (since this likelihood is a subjective mix of the physical probabilities: $C(e|h) = C(e|ha_1)C(a_1|h) + C(e|ha_2)C(a_2|h) + \cdots$ where the a_is are the rival auxiliaries).[17] So defined, the means of calculation remain constant even as both the probabilities of the auxiliaries and the likelihood of the hypothesis simpliciter change (if new auxiliaries appear on the scene) from cycle to cycle.

I call this record the investigation's *evidence structure*. I promised you some rudimentary ethnography of science. Here it is: science does in fact keep track of the evidence structure.[18] That is what the back issues of journals

17. I note in passing that when setting initial priors, the main hypotheses and the auxiliaries might in most cases be treated as statistically independent, further lightening the computational burden. Independence will almost certainly disappear once the evidence begins to arrive (Strevens 2001, note 7).

18. More exactly, it keeps track of "sufficient statistics" about the evidence, that is, enough information about the evidence to determine the likelihoods of the hypotheses on the evidence (Skyrms 1983). Note, however, that what counts as sufficient depends on the family of hypotheses under consideration. The appearance of a new theory (or the reconsideration of an old theory previously considered outlandish) may render what were earlier regarded as sufficient statistics insufficient: it may be that to test the new theory, you need to look at aspects of the evidence above and beyond those that you previously considered important. Assuming that you have not thrown this information away but have merely neglected it, what you must then do is not dissimilar to what is done regularly with authority assessments: you must take into account evidence that is in my sense available but

are for; that is what the Royal Society took as one of its principal tasks; that is, I would argue, the secret function of classical statistics.

The computational load of the suggested system consists in a dramatically holistic recalculation of every probability whenever a new theory is conceived. In practice, the recalculation might not be so holistic: just because the population ecologists come up with a new idea, the high energy physicists should not have to rethink their own epistemic probabilities (and vice versa). But still, it looks like a good deal of work. Too much work?

The recalculation is in fact rather straightforward. Or rather, once the impact of the old evidence on the new theory has been assessed—which may not be straightforward, but which is something that has to be done by anyone's lights—the recalculation of new probabilities for the other, older theories is simple. An updated probability for an old hypothesis h is calculated from the initial priors according to the usual Bayesian rule:

$$\frac{C_1(e|h)}{C_1(e)}C_1(h)$$

where e is your total evidence to date and $C_1(\cdot)$ is the new initial prior probability distribution over both hypotheses and evidence. The mathematics is simple and the values in question are already known to you: you have kept a record of $C_1(h)$, the initial prior for h; you have already calculated the initial prior for the evidence $C_1(e)$ in order to assess the impact of the evidence on your new theory (its value does not depend on the identity of the hypothesis under evaluation); and $C_1(e|h)$ depends only on physical probabilities or entailments recorded in your evidence structure. (When there are auxiliary hypotheses involved, so that there is no value for $C_1(e|h)$ per se in the evidence structure, a separate recalculation will be made for

which you had previously, for perfectly good reasons, ignored. In short, a certain amount of epistemic backtracking is called for, but now with regard to the evidence structure itself. As I remarked earlier, the importance of epistemic backtracking extends beyond the epistemology of scientific authority to any situation in which "bounded rationality" militates against the examination of all available evidence.

every hypothesis/auxiliary combination.) Your recalculation, then, is not at all onerous.[19]

Call the proposed system for dealing with old evidence *recurrent Bayesianism*. Recurrent Bayesianism departs from Bayesian orthodoxy—as any substantive solution to the problem of old evidence must—in two ways. First and more obvious is the feature that gives it its name, the eternal cycle of reconsideration of the priors and reconditionalization. In a sense, the cycle is not deeply unorthodox: conditionalization remains the only mechanism for updating priors, so that your probabilities at any time are the result of setting priors and conditionalizing on the evidence, as according to the traditional Bayesian code.

Second and more controversial is recurrent Bayesianism's dispensing with the "catch-all" hypothesis, or in other words, its failure to represent a scientist's (presumably non-zero) subjective probability that some as-yet unknown theory is correct. It follows from this omission that the subjective probabilities that appear explicitly in the recurrent Bayesian apparatus cannot represent the scientist's actual epistemic state. Additionally, any probabilities to which the catch-all would make a contribution in orthodox Bayesianism—most notably, the prior probability for the evidence $C_1(e)$—must be understood

19. Another way to do the calculation does not invoke the likelihoods in the evidence structure explicitly, but rather relies on the fact that they do not change from cycle to cycle. This method sets the new probability of h equal to

$$\frac{C_1(h)}{C_0(h)}\frac{C_0(e)}{C_1(e)}C(h)$$

where $C_0(\cdot)$ is the old initial prior distribution (the distribution before the initial priors were rearranged to accommodate the new theory), $C_1(\cdot)$ is as before the new initial prior distribution (the distribution after the rearrangement) and $C(h)$ is the current probability for h (the probability in the light of all the evidence, but not the new theory). If priors are "taxed" in proportion to their size to provide the probability assigned to new theories, both ratios in the formula are the same for all hypotheses, so there is a single proportional adjustment to be made to all your epistemic probabilities for the pre-existing hypotheses. It could not be easier. When authority is introduced below, it does get somewhat more complicated, though in many cases the same kind of shortcut can be taken.

within the recurrent Bayesian system not as representing the corresponding degree of belief, for example the scientist's degree of belief in *e*, but rather as representing only that element of the degree of belief due to the known theories. In short, recurrent Bayesianism represents an aspect, but not the totality, of the scientist's epistemic state. Some philosophers want from Bayesianism nothing less than a total epistemology, a complete and self-contained description of everything about a knower's epistemic outlook. Other philosophers see Bayesian confirmation theory as a useful model for some elements of scientific thinking—the more it can model, the better, but its usefulness does not hinge on its being the entire epistemic story. Recurrent Bayesianism will appeal more to the latter, modeling mindset than to the former, totalizing mindset. Whether recurrent Bayesianism can be embedded in some totalizing Bayesian story that explicitly represents a subjective probability for the catch-all hypothesis, I do not know.

4.6 Epistemic Backtracking with Recurrent Bayesianism

Recurrent Bayesianism looks like just the kind of thing needed to deal with the provisionality of assessments of authority-based weight. On the one hand, unlike section 4.1's discount scheme or section 4.4's system of "provisionalities" it provides the resources for real epistemic backtracking in the light of new information. On the other hand, unlike section 4.2's Bayesian tracking or section 4.3's system of second-order probabilities, it inserts new possibilities into the epistemic framework only when they become relevant, thus does not need to quietly but explicitly represent the complete range of epistemically relevant states of affairs ahead of time. Can these promising features be extended to handle authority assessment? Yes; let me show you how.

I will adopt what I called at the end of the previous section the modeling mindset. My aim is to use recurrent Bayesianism to describe a particular aspect of theory confirmation—the bringing to bear of the results published

in journals on scientific theories—not all elements of inductive reasoning in science. The most obvious omission is the absence of any representation of the reasoning that goes into making authority assessments, not least the cost-benefit analyses that tell you how much to invest in such reasoning in the first place. As you will see, this narrowing of focus will have a considerable influence on what, relative to my recurrent Bayesian system, counts as a "prior probability" or as "evidence."

Further, I will simply borrow from the discussion of "second-order non-probability" (section 4.4) the idea that to any first-order probability may be attached a quantity—the probability's "provisionality"—that in some way represents the amount of available evidence taken into account in setting the probability. I will not describe any particular scheme for representing or keeping track of these provisionalities; that task, too, remains on the "to do" list.

I have isolated three places in which authority assessment matters in scientific testing: in determining the initial plausibility of hypotheses, in determining the reliability of evidence, and in determining the trustworthiness of auxiliary hypotheses. Take recurrent Bayesianism and represent the influence of authority at these three points as follows.

Plausibility of Hypotheses Consider your initial prior probabilities for some set of competing hypotheses in the recurrent Bayesian system. As I conceive it, the system allows that these priors may be based in great part on expert opinion as to the hypotheses' plausibility. Thus, they are not "prior" in the totalizing Bayesian's sense; they are not formed in an empirical void. Rather, they take into account what the experts think, in advance of their assessing the experimental evidence—an empirical fact, but not an *experimental* fact. The priors in my recurrent Bayesianism are prior only to experiment.

In assessing the expert's opinion of a hypothesis's initial plausibility, you

may for all the reasons given above decide not to seek out and weigh all available evidence. Accordingly, attach a "provisionality" (the same sort of second-order quantity examined in section 4.4 above) to each prior to represent the amount of available evidence taken into account.

Now suppose that some hypothesis attains a new-found importance that merits a reexamination of the authority assessments on which its initial prior probability is based. You take into account more of the available evidence, and perhaps uncover new evidence, that bears on these assessments—evidence concerning the reliability of the experts in question, the degree to which they arrived at their conclusions independently, and so on. As a result, let me suppose, you reconsider your initial prior for the hypothesis, arriving at a new value. What to do about this new value? Simple: recalculate your probabilities just as you do when you come up with a new theory. (You will first have to normalize your initial prior distribution, by redistributing probability among the other initial priors to make up for the change in the prior under examination.)

Experimental Outcomes Consider next uncertainty about the evidence. I will suppose that you represent the possibility that an experiment has gone wrong by assigning a posterior probability of less than one to the result claimed by the experimenter. On publication of the results, then, your probability for the relevant evidence statement e goes up, but not all the way to one.[20]

20. This is an oversimplification. An experimental report describes not only an outcome, but the set of initial and background conditions in the context of which the outcome was produced. Such a report has two parts, then, the conditions c and the outcome e; the form of the likelihood of a hypothesis h on the evidence is $C(e|hc)$. When an experimental result is assigned a posterior probability of less than one, it is often the status of c that is in question; for this reason, it is misleading to treat the impact of the experiment as a simple Jeffrey conditionalization on e, as I do in the next paragraph. To put it another way: when you mistrust an experimenter's reported result e, it is often because you suspect that because of some flaw in the setup, e's occurrence provides no relevant information, rather than because you think that the setup was fine but it was in fact $\neg e$ that occurred. These issues are,

In the recurrent Bayesian system I advocate here, the change in *e*'s probability is exogenous. That is, the system does not represent any of the reasoning that goes into determining the post-publication probability of *e*; in particular, it does not represent your reasoning about the reliability of the experimenter. The experimental proposition *e* therefore plays the role of the evidence in the Bayesian sense: it undergoes a change "from the outside" which then triggers an episode of conditionalization—or rather, because the posterior probability of *e* is less than one, Jeffrey conditionalization (Jeffrey 1983). A more totalizing Bayesian treatment would suppose that, on publication of the results, the probability of *something* went up to one ("such and such a sentence appeared in such and such a journal"), and that the rise in the probability of *e* is the result of conditionalizing on that something. My non-standard handling makes the Bayesian's "evidence" and the scientist's "evidence" one and the same set of facts.[21]

In assessing the reliability of an experimenter, you typically do not take all available evidence into account; attach a provisionality, then, to every experimental claim, representing the amount of available evidence taken into account when making the authority assessments that went into determining the posterior probability of the claim.

Suppose that a change in scientific priorities motivates a closer look at some controversial experiment. More evidence about the experimenter's reliability, and about the reliability of those who assess the experimenter's reliability, and so on, is collected. The posterior probability for *e* is adjusted as a result (and its provisionality reduced). What next? Recalculate your probabilities for the hypotheses to which *e* is relevant using the new posterior for *e*, as recurrent Bayesianism instructs.

---

however, orthogonal to the principal concerns of this paper; I will ignore them in the main text.

21. For Jeffrey's own treatment of "unreliable testimony" using Jeffrey conditionalization, see Jeffrey (1987).

Auxiliary Hypotheses Third, consider uncertainty about the auxiliary hypotheses that play a part in determining the likelihood of some hypothesis on some piece of evidence. Again, represent the quality of your evidence for the relevant authority assessment by a provisionality; again, adjust the probabilities of auxiliary hypotheses when necessary as changing cost-benefit considerations dictate; again, when the probabilities change, recalculate.

Recalculation How demanding are the recalculations I have proposed? A change in the initial prior of a hypothesis is very easy to accommodate: since the posterior probabilities are linear functions of the initial priors, they will change in proportion to those priors. If a reconsideration of expert opinion causes the initial prior of a hypothesis to double, for example, its posterior will also double (see note 19).

A change in the posterior of a piece of evidence calls for a slightly more difficult recalculation, since that evidence will be differently relevant to different hypotheses. But again, it is ultimately a matter of simple proportionality. The probability of a hypothesis h is impacted by an experimental result e in proportion to the probability of e, the likelihood of h on e (and now, I should add, on $\neg e$, though see note 20), and the prior probability of e. Provided that a record of likelihoods is kept, this impact can be adjusted accordingly as the probability of e changes.

Finally, a change in the posterior of an auxiliary is dealt with in much the same way. In this case, however, it is not enough to have a record of the likelihood of the hypothesis on the evidence, since it is this very value that is affected when the probability of the auxiliary is altered. What you need is rather the information needed to calculate the likelihood. This comes in three parts: the contents of the main hypothesis itself, the content of the auxiliaries that are used to bring the main hypothesis to bear on—that is, to declare a probability for—a particular empirical outcome, and the probabilities of the auxiliaries. The first two parts are (by design) found in the evidence structure

and so are by assumption readily available; the third is what changes as a result of the reallocation of probability among the auxiliaries. The rest is simply a matter of multiplication.[22]

Note that you might revise your authority-based probability for an auxiliary hypothesis for two reasons. First, as with the other authority-based probabilities considered so far, you might for various reasons revise your opinions about the reliability of the different scientists who are experts in the auxiliary domain. But second, the experts might themselves, in the light of new evidence, change their minds about the plausibility of the relevant auxiliaries. By this latter route, recurrent Bayesianism allows you to offload the work of tracking the evidence for some hypotheses to experts. As the experts learn more, you adjust your own epistemic state to take into account their new knowledge, using exactly the same machinery as when you learn more about the reliability of the experts. (This distinction between two ways of using authority-based probability constitutes an alternative to the proximal/distal dichotomy set out at the beginning of this paper.)

* * *

Let me take stock. Recurrent Bayesian epistemic backtracking is, I propose, a golden mean that lies between, on the one hand, the elaborate but unrealistic Bayesian tracking and second-order probability systems—systems that allow backtracking but at exorbitant cognitive expense—and on the other hand, systems that are simple and straightforward to implement but that do not provide the resources for genuine backtracking. Although the recurrent backtracking system is developed in a Bayesian framework, I hope that it is

22. Because the recurrent system breaks out the likelihoods of each hypothesis/auxiliary package separately, as explained in the previous section, the case in which priors for the auxiliaries change is in fact structurally identical to the case in which the priors for the main hypotheses change; it is more intuitive, however, to think of the recalculation as having two steps, one in which new likelihoods for the hypotheses simpliciter are calculated, and one in which these likelihoods are used to calculate new probabilities for the hypotheses themselves. (And there is of course a third step: calculating new probabilities for the auxiliaries in the light of the evidence.)

clear enough that its basic structure can be implemented in non-Bayesian terms.[23]

Two items of unfinished business. First, observe that the recurrent Bayesian system provides a way to handle what I called in section 3.1 the determination problem, that is, the problem arising from the fact that in assessing an authority's reliability at some task, a scientist is often unclear in their own mind as to the identity of the task in question. This epistemic imperfection (if that is the right word, when the unclarity is entirely justified on cost-benefit grounds) is not a violation of the principle of total evidence, but rather a violation of the injunction to *be precise about your hypothesis space*. Suppose that a change in your priorities motivates a precisification of your hypothesis space, which in turn causes a change in your authority-based probabilities. Backtracking works all the same: whatever your reasons for wishing to adjust your authority-based epistemic probabilities, backtracking provides you with the means to change their values, and the means to recalculate everything that depends on those values.[24]

Second, do you need to keep track of the particulars of all the available evidence you have taken into account in making an authority assessment, so that when you go back to reconditionalize, you can recognize what evidence is already factored into your initial prior probabilities and what has yet to be incorporated? Not necessarily; if backtracking is relatively rare, it might be easier simply to start all over again when backtracking; what you expend in footstep-retracing you more than save in record-keeping.

23. On some approaches to confirmation, there is no need for initial prior probabilities, thus no need to reconsider them when backtracking. They may, accordingly, be omitted from the backtracking structure.

24. Backtracking also provides a way of coping, *within* the epistemology of authority assessment, with a change in the space of hypotheses to which initial priors are assigned. But such issues lie outside the scope of the present paper.

5. Ethnographic Questions

What reason is there to think that real science implements something like my kind of epistemic backtracking? What reason is there to think that scientists keep a record of the information needed for backtracking, and if they do, that they use it to backtrack in the recurrent Bayesian way?

Recurrent Bayesian backtracking requires scientists to represent the following epistemic probabilities and facts, for any set of rival hypotheses of interest:

1. The initial plausibility of the hypotheses, that is, their initial priors.

2. The relevant evidence accrued.

3. The reliability of the evidence, that is, the probability that nothing went wrong in the experiment or observation generating the evidence.

4. The structure of the likelihood calculation—most notably, the physical probabilities bestowed on the evidence by different hypothesis/auxiliary packages.

5. The probabilities of the auxiliary hypotheses invoked in the likelihood calculation.

Is there any evidence that they do represent this information?

Divide these elements into two groups. The first group comprises (2) and (4), the catalog of evidence and the structure of the likelihood calculations for the evidence—what I have called the *evidence structure*. As I remarked above, there is no question that scientists keep track of this information; it is what scientists publish, when they publish.

The evidence structure is a public part of science. The second group—(1), (3), and (5), or the initial priors, the estimates of the reliability of the evidence, and the estimates of the reliability of the auxiliary hypotheses—are normally private; they are not recorded in the journals. But it is easy to find informal

discussion of these matters. The question of the reliability of evidence is a particularly sensitive one, of course, potentially impugning as it does a colleague's competence, but it is openly debated "over beer" (a technical term among scientific ethnographers); for some examples, see Collins (1975, 214–215). In any case, it is hard to see, even without backtracking, how science could possibly proceed without such estimates.

Suppose, then, that scientists have the information they need in order to backtrack. Do they do it? I have no direct evidence for an affirmative answer, but with the information at their fingertips, and with the need for some sort of backtracking being so clear—if what I have written above about the role of authority in science is correct—it would seem perverse for scientists not to backtrack on occasion. Whether they backtrack in precisely the way I have suggested here (or close enough that recurrent Bayesianism can serve as a model for real-life backtracking) is a question I leave to another time.

And outside science? In everyday life? Let me provide questions rather than answers. To what extent do the epistemic probabilities (or the equivalent) in our everyday epistemology depend on authority? (I will answer that one: quite a bit, many philosophers believe.) To what extent are the relevant authority assessments provisional? Do we engage in cost-benefit analyses to determine how much of the available evidence to take into account in assessing authority? If so, how often do the terms of these analyses change over time, motivating us to gather more evidence, including more of the evidence that was available but that went unused in the initial assessment? And when an assessment changes in the light of a more extensive examination of the evidence, how do we deal with the change? Through epistemic backtracking?

6. The Public and the Private in Science

It is peculiar that the public record of science describes only the evidence structure, when clearly so much else is epistemically important. Why, in

particular, no posterior probabilities—even approximate probabilities—for hypotheses? This silence has some serious practical disadvantages. In particular, when making informed decisions about public policy, such as decisions as to how (if at all) global warming should be tackled, policy-makers cannot simply consult the journals to get the current consensus estimate of the way that the evidence is pointing. They must instead elicit opinions from a selection of experts, with the attendant problems of selection bias, observer effect, and so on.

There are a number of possible explanations for the public/private divide. Perhaps scientists are convinced of the official frequentist ideology of classical statistics, on which it makes no sense to attach probabilities to hypotheses (though this strikes me as a post hoc justification for a practice otherwise motivated). Perhaps the evidence structure plays a special inductive role in science (Glymour 1980; Strevens manuscript). Perhaps some norm of science enjoins scientists to publish only probabilities (and other epistemically relevant quantities) on which there can be robust intersubjective agreement (Strevens 2009). The discussion in this paper suggests another possibility, related to the last: what is published is that part of scientific epistemology that does not rely on authority.

Consider an epistemically relevant datum that derives in part from scientific authority, such as a consensus epistemic probability that some experimental result—the cold fusion data, for example—is veridical. What is it about the datum's dependence on authority that discourages its publication? No doubt it is partly a matter of uncertainty concerning the existence of the consensus in the first place. But also a problem is the datum's provisionality. What is published in the journals becomes a part of science's permanent record. A provisional estimate is subject to change. Further, it is subject to change not only because new evidence may arrive, but because some currently available evidence has been ignored—you cannot, then, publish the datum even with a rider saying that it reflects the state of knowledge at the

time of publication. Provisionally determined values are for these reasons unsuitable for the record.

The concern with the permanence of the record creates a curious duality, a dualism even, in the epistemology of science. The public epistemology of science—or, you might say, the *published* epistemology of science—is built around what I have called the evidence structure. "Frequentist" philosophers of statistics (that is, defenders of classical statistics) and their fellow travelers, in particular likelihood theorists, have made great efforts to argue that the evidence structure provides all the epistemology that science needs (Edwards 1972; Mayo 1996).

But the evidence structure omits every aspect of scientific epistemology that depends in part on authority assessment, and authority assessment is essential to the acquisition of scientific knowledge. Thus there exists alongside the public epistemology of science a private epistemology, an epistemology that makes room for authority, for provisionality, for epistemic backtracking.

To some extent the private epistemology may be incarnated differently in the mind of every scientist. Provided that scientists pay due heed to authority, however, their epistemic states will be in many ways coordinated; there will be disagreements between scientists, but they will tend to concern the few issues on which those scientists are (or consider) themselves the authorities. Thus there is something that constitutes a partial scientific consensus above and beyond the consensus on the evidence structure, even if it is deliberately hidden from view.

Under what circumstances can the private scientific epistemology be observed at work? Not in the textbooks; a hypothesis makes its appearance in the canon only once it has sloughed off all its epistemic properties such as "well-evidenced" and "justifiable" to become simply "true"—a matter of fact. Always "over beer," if you know where the scientists go to unwind. But the private epistemology can also be observed whenever important decisions turn on the truth or otherwise of hypotheses that are as yet under investigation,

that is, whenever scientists must, for the greater good of humankind or its local instantiation, expose their epistemic probabilities to the scrutiny of the makers of public policy. Then, summoned by various techniques for epistemic extraction—such as the Delphi method (Linstone and Turoff 1975), or prediction markets, or the promise of television appearances or expert witness fees—epistemic probabilities, still writhing with provisionality, worm their way toward daylight.

Acknowledgments

Thanks to Keith DeRose and Timothy Williamson for very helpful feedback.

References

Carnap, R. (1950). *Logical Foundations of Probability*. University of Chicago Press, Chicago.

Christensen, D. (2007). Epistemology of disagreement: The good news. *Philosophy of Science* 116:187–217.

Coady, C. A. J. (1995). *Testimony: A Philosophical Study*. Oxford University Press, Oxford.

Collins, H. M. (1975). The seven sexes: A study in the sociology of a phenomenon, or the replication of experiments in physics. *Sociology* 9:205–224.

De Cooman, G. and P. Walley. (2002). A possibilistic hierarchical model for behavior under uncertainty. *Theory and Decision* 52:327–374.

Earman, J. (ed.). (1983). *Testing Scientific Theories*, volume 10 of *Minnesota Studies in the Philosophy of Science*. University of Minnesota Press, Minneapolis.

Edwards, A. W. F. (1972). *Likelihood: An Account of the Statistical Concept of Likelihood and Its Application to Scientific Inference*. Cambridge University Press, Cambridge.

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics* 75:643–669.

Foley, R. (1994). Egoism in epistemology. In F. Schmitt (ed.), *Socializing Epistemology*. Rowman & Littlefield, Lanham, MD.

Garber, D. (1983). Old evidence and logical omniscience in Bayesian confirmation theory. In Earman (1983).

Glymour, C. (1980). *Theory and Evidence*. Princeton University Press, Princeton, NJ.

Goldman, A. I. (1999). *Knowledge in a Social World*. Oxford University Press, Oxford.

Good, I. J. (1980). Some history of the hierarchical Bayesian methodology. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (eds.), *Bayesian Statistics*, volume 1. University of Valencia Press, Valencia, Spain.

Hardwig, J. (1985). Epistemic dependence. *Journal of Philosophy* 82:335–349.

Howson, C. (1984). Bayesianism and support by novel facts. *British Journal for the Philosophy of Science* 35:245–251.

Hull, D. (1988). *Science as a Process*. University of Chicago Press, Chicago.

Jeffrey, R. C. (1983). *The Logic of Decision*. Second edition. University of Chicago Press, Chicago.

———. (1987). Alias Smith and Jones: The testimony of the senses. *Erkenntnnis* 26:391–399.

Kelly, T. (2005). The epistemic significance of disagreement. In T. S. Gendler and J. Hawthorne (eds.), *Oxford Studies in Epistemology*, volume 1. Oxford University Press, Oxford.

———. (Forthcoming). Peer disagreement and higher order evidence. In R. Feldman and T. Warfield (eds.), *Disagreement*. Oxford University Press, Oxford.

Keynes, J. M. (1921). *A Treatise on Probability*. Macmillan, London.

Kitcher, P. (1993). *The Advancement of Science*. Oxford University Press, Oxford.

Lackey, J. and E. Sosa (eds.). (2006). *The Epistemology of Testimony*. Oxford University Press, Oxford.

Lehrer, K. and C. Wagner. (1981). *Rational Consensus in Science and Society*. D. Reidel, Dordrecht.

Lewis, D. (1980). A subjectivist's guide to objective chance. In R. C. Jeffrey (ed.), *Studies in Inductive Logic and Probability*, volume 2. University of California Press, Berkeley, CA.

Linstone, H. A. and M. Turoff (eds.). (1975). *The Delphi Method: Techniques and Applications*. Addison-Wesley, Reading, MA.

Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*. University of Chicago Press, Chicago.

Polanyi, M. (1958). *Personal Knowledge: Towards a Post-Critical Philosophy*. University of Chicago Press, Chicago.

Schervish, M. J. (1989). A general method for comparing probability assessors. *Annals of Statistics* 17:1856–1879.

Shapin, S. (1994). *A Social History of Truth: Civility and Science in Seventeenth-Century England*. University of Chicago Press, Chicago.

Skyrms, B. (1980). Higher order degrees of belief. In D. H. Mellor (ed.), *Prospects for Pragmatism : Essays in Memory of F. P. Ramsey*. Cambridge University Press, Cambridge.

———. (1983). Three ways to give a probability assignment a memory. In Earman (1983).

Strevens, M. (1995). A closer look at the 'New' Principle. *British Journal for the Philosophy of Science* 46:545–561.

———. (2001). The Bayesian treatment of auxiliary hypotheses. *British Journal for the Philosophy of Science* 52:515–538.

———. (2009). Objective evidence and absence. *Philosophical Studies* 143:91–100.

———. (Manuscript). What is empirical testing?

Williamson, T. (2000). *Knowledge and Its Limits*. Oxford University Press, Oxford.

Zollman, K. J. S. (2007). The communication structure of epistemic communities. *Philosophy of Science* 74:574–587.

———. (Forthcoming). The epistemic benefit of transient diversity. *Erkenntnis*.