Excursion 4 Objectivity and Auditing

Itinerary

Tour I The Myth of "The Myth of Objectivity"

> Objectivity in statistics, as in science more generally, is a matter of both aims and methods. Objective science, in our view, aims to find out what is the case as regards aspects of the world [that hold] independently of our beliefs, biases and interests; thus objective methods aim for the critical control of inferences and hypotheses, constraining them by evidence and checks of error. (Cox and Mayo 2010, p. 276)

Whenever you come up against blanket slogans such as "no methods are objective" or "all methods are equally objective and subjective" it is a good guess that the problem is being trivialized into oblivion. Yes, there are judgments, disagreements, and values in any human activity, which alone makes it too trivial an observation to distinguish among very different ways that threats of bias and unwarranted inferences may be controlled. Is the objectivity–subjectivity distinction really toothless, as many will have you believe? I say no. I know it's a meme promulgated by statistical high priests, but you agreed, did you not, to use a bit of *chutzpah* on this excursion? Besides, cavalier attitudes toward objectivity are at odds with even more widely endorsed grass roots movements to promote replication, reproducibility, and to come clean on a number of sources behind illicit results: multiple testing, cherry picking, failed assumptions, researcher latitude, publication bias and so on. The moves to take back science are rooted in the supposition that we can more objectively scrutinize results – even if it's only to point out those that are BENT. The fact that these terms are used equivocally should not be taken as grounds to oust them but rather to engage in the difficult work of identifying what there is in "objectivity" that we won't give up, and shouldn't.

The Key Is Getting Pushback! While knowledge gaps leave plenty of room for biases, arbitrariness, and wishful thinking, we regularly come up against data that thwart our expectations and disagree with the predictions we try to foist upon the world. We get pushback! This supplies objective constraints on which our critical capacity is built. Our ability to recognize when data fail to match anticipations affords the opportunity to systematically improve our orientation. Explicit attention needs to be paid to communicating results to set the stage for others to check, debate, and extend the inferences reached.

Which conclusions are likely to stand up? Where do the weakest parts remain? Don't let anyone say you can't hold them to an objective account.

Excursion 2, Tour II led us from a Popperian tribe to a workable demarcation for scientific inquiry. That will serve as our guide now for scrutinizing the myth of the myth of objectivity. First, good sciences put claims to the test of refutation, and must be able to embark on an inquiry to pin down the sources of any apparent effects. Second, refuted claims aren't held on to in the face of anomalies and failed replications; they are treated as refuted in further work (at least provisionally); well-corroborated claims are used to build on theory or method: science is not just stamp collecting. The good scientist deliberately arranges inquiries so as to capitalize on pushback, on effects that will not go away, on strategies to get errors to ramify quickly and force us to pay attention to them. The ability to register how hunting, optional stopping, and cherry picking alter their error-probing capacities is a crucial part of a method's objectivity. In statistical design, day-to-day tricks of the trade to combat bias are consciously amplified and made systematic. It is not because of a "disinterested stance" that we invent such methods; it is that we, quite competitively and self-interestedly, want our theories to succeed in the market place of ideas.

Admittedly, that desire won't suffice to incentivize objective scrutiny if you can do just as well producing junk. Successful scrutiny is very different from success at grants, getting publications and honors. That is why the reward structure of science is so often blamed nowadays. New incentives, gold stars and badges for sharing data and for resisting the urge to cut corners are being adopted in some fields. Fortunately, for me, our travels will bypass lands of policy recommendations, where I have no special expertise. I will stop at the perimeters of scrutiny of methods which at least provide us citizen scientists armor against being misled. Still, if the allure of carrots has grown stronger than the sticks, we need stronger sticks.

Problems of objectivity in statistical inference are deeply intertwined with a jungle of philosophical problems, in particular with questions about what objectivity demands, and disagreements about "objective versus subjective" probability. On to the jungle!

4.1 Dirty Hands: Statistical Inference Is Sullied with Discretionary Choices

> If all flesh is grass, kings and cardinals are surely grass, but so is everyone else and we have not learned much about kings as opposed to peasants. (Hacking 1965, p. 211)

Trivial platitudes can appear as convincingly strong arguments that everything is subjective. Take this one: No human learning is pure so anyone who demands objective scrutiny is being unrealistic and demanding immaculate inference. This is an instance of Hacking's "all flesh is grass." In fact, Hacking is alluding to the subjective Bayesian de Finetti (who "denies the very existence of the physical property [of] chance" (ibid.)). My one-time colleague, I. J. Good, used to poke fun at the frequentist as "denying he uses any judgments!" Let's admit right up front that every sentence can be prefaced with "agent x judges that," and not sweep it under the carpet (SUTC) as Good (1976) alleges. Since that can be done for any statement, it cannot be relevant for making the distinctions in which we are interested, and we know can be made, between warranted or well-tested claims and those so poorly probed as to be BENT. You'd be surprised how far into the thicket you can cut your way by brandishing this blade alone.

It is often urged that, however much we may aim at objective constraints, we can never have clean hands, free of the influence of beliefs and interests. We invariably sully methods of inquiry by the entry of background beliefs and personal judgments in their specification and interpretation. The real issue is not that a human is doing the measuring; the issue is whether that which is being measured is something we can reliably use to solve some problem of inquiry. An inference done by machine, untouched by human hands, wouldn't make it objective in any interesting sense. There are three distinct requirements for an objective procedure of inquiry:

1. *Relevance:* It should be relevant to learning about what is being measured; having an uncontroversial way to measure something is not enough to make it relevant to solving a knowledge-based problem of inquiry.
2. *Reliably capable:* It should not routinely declare the problem solved when it is not (or solved incorrectly); it should be capable of controlling reports of erroneous solutions to problems with reliability.
3. *Able to learn from error:* If the problem is not solved (or poorly solved) at a given point, the method should set the stage for pinpointing why.

Yes, there are numerous choices in collecting, analyzing, modeling, and drawing inferences from data, and there is often disagreement about how they should be made, and about their relevance for scientific claims. Why suppose that this introduces subjectivity into an account, or worse, means that all accounts are in the same boat as regards subjective factors? It need not, and they are not. An account of inference shows itself to be objective precisely in how it steps up to the plate in handling potential threats to objectivity.

Dirty Hands Argument. To give these arguments a run for their money, we should try to see why they look so plausible. One route is to view the reasoning as follows:

1. A variety of human judgments go into specifying experiments, tests, and models.
2. Because there is latitude and discretion in these specifications, which may reflect a host of background beliefs and aims, they are "subjective."
3. Whether data are taken as evidence for a statistical hypothesis or model depends on these subjective methodological choices.
4. Therefore, statistical methods and inferences are invariably subjective, if only in part.

The mistake is to suppose we are incapable of critically scrutinizing how discretionary choices influence conclusions. It is true, for example, that choosing a very insensitive test for detecting a risk δ' will give the test low probability of detecting such discrepancies even if they exist. Yet I'm not precluded from objectively determining this. Setting up a test with low power against δ' might be a product of your desire not to find an effect for economic reasons, of insufficient funds to collect a larger sample, or of the inadvertent choice of a bureaucrat. Or ethical concerns may have entered. But our critical evaluation of what the resulting data do and do not indicate need not itself be a matter of economics, ethics, or what have you.

Idols of Objectivity. I sympathize with disgruntlement about phony objectivity and false trappings of objectivity. They grow out of one or another philosophical conception about what objectivity requires – even though you will almost surely not see them described that way. It's the curse of logical positivism, but also its offshoots in post-positivisms. If it's thought objectivity is limited to direct observations (whatever they are) plus mathematics and logic, as the typical positivist, then it's no surprise to wind up worshiping what Gigerenzer and Marewski (2015) call "the idol of a universal method." Such a method is to supply a formal, ideally mechanical, rule to process statements of observations and hypotheses – translated into a neutral observation language. Can we translate Newtonian forces and Einsteinian curved spacetime into a shared observation language? The post-positivists, rightly, said no. Yet giving up on logics of induction and theory-neutral languages, they did not deny these were demanded by objectivity. They only decided that they were unobtainable. Genuine objectivity goes by the board, replaced by various

stripes of relativism and constructivism, as well as more extreme forms of anarchism and postmodernism.[1]

From the perspective of one who has bought the view that objectivity is limited to math, logic, and fairly direct observations (the dial now points to 7), methods that go beyond these appear "just as" subjective as another. They may augment their rather thin gruel with an objectivity arising from social or political negotiation, or a type of consensus, but that's to give away the goat far too soon. The result is to relax the core stipulations of scientific objectivity. To be clear: There are authentic problems that threaten objectivity. Let's not allow outdated philosophical accounts to induce us into giving it up.

What about the fact that different methods yield different inferences, for example that Richard Royall won't infer the composite $\mu > 0.2$ while N-P testers will? I have no trouble understanding why, if you define inference as comparative likelihoods, the results disagree with error statistical tests. Running different analyses on the same data can be the best way to unearth flaws. However, objectivity is an important criterion in appraising such rival statistical accounts.

Objectivity and Observation. In facing objectivity skeptics, you might remind them of parallels between learning from statistical experiments and learning from observations in general. The problem in objectively interpreting observations is that observations are always relative to the particular instrument or observation scheme employed. But we are often aware not only of the fact that observation schemes influence what we observe but also of how: How much noise are they likely to introduce? How might we subtract it out?

The result of a statistical method need only (and should only) be partly determined by the specifications of a given method (e.g., the cut-off for statistical significance); it is also determined by the underlying scientific phenomenon, as modeled. What enables objective learning is the possibility of taking into account *how* test specifications color results as we intervene in phenomena of interest. Don't buy the supposition that the word "arbitrary" always belongs in front of "convention." That my weight shows up as k pounds is a convention in the USA. Still, *given the convention*, the readout of k pounds is a matter of how much I weigh. I cannot simply ignore the additional weight as due to an arbitrary convention, even if I wanted to.

[1] See Larry Laudan's (1996) ingenious analysis of how much the positivists and post-positivists share in "The Sins of the Fathers."

How Well Have You Probed *H* versus How Strongly Do (or Should) You Believe It?

When Albert Einstein was asked "What if Eddington had not found evidence of the deflection effect?", Einstein famously replied, "Then I would feel sorry for the dear Lord; the theory is correct." Some might represent this using subjective Bayesian resources. Einstein had a strong prior conviction in GTR – a negative result might have moved his belief down a bit but it would still be plenty high. Such a reconstruction may be found useful. If we try to cash it out as a formal probability, it isn't so easy. Did he assign high prior to the deflection effect being 1.75″, or also to the underlying theoretical picture of curved spacetime (which is really the basis of his belief)? A formal probability assignment works better for individual events than for assessing full-blown theories, but let us assume that it could be done. What matters is that Einstein would also have known the deflection hypothesis had not been well probed, that is, it had not yet passed a severe test in 1919. An objective account of statistics needs to distinguish how probable (believable, plausible, supported) a claim is from how well it has been probed. This remains true whether the focus is on a given set of data, several sets, or, given everything I know, what I called "big picture" inference.

Having distinguished our aim – appraising how stringently and responsibly probed a claim *H* is by the results of a given inquiry – from that of determining *H*'s plausibility or belief-worthiness, it's easy to allow that different methodologies and criteria are called for in pursuing these two goals. Recall the root of *probare* is to demonstrate or show.

Some argue that "discretionary choices" in tests, which Neyman himself tended to call "subjective," lead us to subjective probabilities. A weak version goes: since you can't avoid discretionary choices in getting the data and model, how can you complain about subjective degrees of belief in the resulting inference? This is weaker than arguing you must use subjective probabilities; it argues merely that doing so *is no worse than* discretion. It still misses the point.

First, as we saw in exposing the "dirty hands" argument, even if discretionary judgments can introduce subjectivity, they need not. Second, not all discretionary judgments are in the same boat when it comes to being open to severe testing of their own. E. Pearson imagines he

> might quote at intervals widely different Bayesian probabilities for the same set of states, simply because I should be attempting what would be for me impossible and resorting to guesswork. It is difficult to see how the matter could be put to experimental test. (Pearson 1962, pp. 278–9)

A stronger version of the argument goes on a slippery slope from the premise of discretion in data generation and modeling to the conclusion: statistical inference is a matter of subjective belief. How does that work? One variant involves a subtle slide from "our models are merely objects of belief," to "statistical inference is a matter of degrees of belief." From there it's a short step to "statistical inference is a matter of subjective probability." It is one thing to allow talk of our models as objects of belief and quite another to maintain that our task is to model beliefs.

This is one of those philosophical puzzles of language that might set some people's eyes rolling. If I believe in the deflection effect then that effect is the object of my belief, but only in the sense that my belief is about said effect. Yet if I'm inquiring into the deflection effect, I'm not inquiring into beliefs about the effect. The philosopher of science Clark Glymour (2010, p. 335) calls this a shift from phenomena (content) to *epiphenomena* (degrees of belief). Popper argues that the key confusion all along was sliding from the degree of the rationality (or warrantedness) of a belief, to the degree of rational belief (1959, p. 407).

Or take subjectivist Frank Lad. To him,

> . . . so-called 'statistical models' are not real entities that merit being estimated. To the extent that models mean anything, they are models of someone's (some group's) considered uncertain opinion about observable quantities. (Lad 2006, p. 443)

Notice the slide from uncertainty or partial knowledge of quantities in models, to models being *models of opinions*. I'm not saying Lad is making a linguistic error. He appears instead to embrace a positivist philosophy of someone like Bruno de Finetti. De Finetti denies we can put probabilities on general claims because we couldn't settle bets on them. If it's also maintained that scientific inference takes the form of a subjective degree of belief, then we cannot infer general hypotheses – such as statistical hypotheses. Are we to exclude them from science as so much meaningless metaphysics?

When current-day probabilists echo such stances, it's a good bet they would react with horror at the underlying logical positivist philosophy. So how do you cut to the chase without sinking into a philosophical swamp? You might ask: Are you saying statistical models are just models of beliefs and opinions? They are bound to say no. So press on and ask: Are you saying they are mere approximations, and we hold fallible beliefs and opinions about them? They're likely to agree. But the error statistician holds this as well!

What's Being Measured versus My Ability to Test It. You will sometimes hear it claimed that anyone who says their probability assignments to hypotheses are subjective must also call the use of any model subjective because it too

is based on my choice of specifications. It's important not to confuse two notions of subjective. The first concerns what's being measured, and for the Bayesian, at least the subjective Bayesian, probability represents a subject's strength of belief. The second sense of subjective concerns whether the measurement is checkable or testable. Nor does latitude for disagreement entail untestability. An intriguing analysis of objectivity and subjectivity in statistics is Gelman and Hennig (2017).

4.2 Embrace Your Subjectivity

The classical position of the subjective Bayesian aims at inner coherence or consistency rather than truth or correctness. Take Dennis Lindley:

> I am often asked if the method gives the *right* answer: or, more particularly, how do you know if you have got the *right* prior. My reply is that I don't know what is meant by 'right' in this context. The Bayesian theory is about *coherence*, not about right or wrong. (Lindley 1976, p. 359)

There's no reason to suppose there is a correct degree of belief to hold. For Lindley, $\Pr(H|\boldsymbol{x})$ "is your belief about [H] when you know [$\boldsymbol{x}$]" (Lindley 2000, p. 302, substituting Pr for P; H for A and $\boldsymbol{x}$ for B). My opinions are my opinions and your opinions are yours. How do I criticize your prior degrees of belief? As Savage said, "[T]he Bayesian outlook reinstates opinion in statistics – in the guise of the personal probabilities of events . . ." (Savage 1961, p. 577). Or again, "The concept of personal probability . . . seems to those of us who have worked with it an excellent model for the concept of opinion" (ibid., pp. 581–2).[2] That might be so, but what if we are not trying to model opinions, but instead insist on meeting requirements for objective scrutiny? For these goals, inner coherence or consistency among your beliefs is not enough. One can be consistently wrong, as everyone knows (or should know).

If you're facing a radical skeptic of all knowledge, a radical relativist, postmodernist, social-constructivist, or anarchist, there may be limited room to maneuver. The position may be the result of a desire to shock or be camp (as Feyerabend or Foucault) or give voice to political interests. The position may be mixed with, or at least dressed in the clothes of, philosophy: We are locked in a world of appearances seeing mere shadows of an "observer independent reality." Our bold activist learner, who imaginatively creates abstract models that give him pushback, terrifies them. Calling it unholy metaphysics may actually reflect their inability to do the math.

[2] I will not distinguish personalists and subjectivists, even though I realize there is a history of distinct terms.

Progress of Philosophy

To the error statistician, radical skepticism is a distraction from the pragmatic goal of understanding how we do manage to learn, and finding out how we can do it better. Philosophy does make progress. Logical positivism was developed and embraced when Einstein's theory rocked the Newtonian worldview. Down with metaphysics! All must be verifiable by observation. But there are no pure observations, no theory-neutral observational languages, no purely formal rules of confirmation holding between any statements of observation and hypotheses. Popper sees probabilism as a holdover from a watered down verificationism, "... under the influence of the mistaken view that science, unable to attain certainty, must aim at a kind of '*Ersatz*' – at the highest attainable probability" (Popper 1959, p. 398 (Appendix IX)). Even in the face of the "statistical crisis in science," by and large, scientists aren't running around terrified that our cherished theories of physics will prove wrong: they expect even the best ones are incomplete, and several rival metaphysics thrive simultaneously. In genetics, we have learned to cut, delete, and replace genes in human cells with the new CRISPR technique discovered by Jennifer Doudna and Emmanuelle Charpentier (2014). The picture of the knower limited by naked observations no longer has any purchase, if it ever did.

Some view the Big Data revolution, with its focus on correlations rather than causes, as a kind of return to theory-free neopositivism. Theory freedom and black-box modeling might work well for predicting what color website button is most likely to get me to click. AI has had great successes. We've also been learning how theory-free prediction techniques come up short when it comes to scientific understanding.

Loss and Cost Functions

The fact that we have interests, and that costs and values may color our interpretation of data, does not mean they should be part of the scientific interpretation of data. Frequent critics of statistical significance tests, economists Stephen Ziliak and Deirdre McCloskey, declare, in relation to me, that "a notion of a severe test without a notion of a loss function is a diversion from the main job of science" (2008a, p. 147). It's unclear if this is meant as a vote for a N-P type of balancing of error probabilities, or for a full-blown decision theoretic account. If it is the latter, with subjective prior probabilities in hypotheses, we should ask: Whose losses? Whose priors? The drug company? The patient? They may lie hidden in impressive Big Data algorithms as Cathy O'Neil (2016) argues.

Remember that a severity appraisal is always in relation to a question or problem, and that problem could be a decision, within a scientific inquiry or

wholly personal. In the land of science, we'd worry that to incorporate into an inference on genomic signatures, say, your expected windfall from patenting it would let it pass without a severe probe. So if that is what they mean, I disagree, and so should anyone interested in blocking flagrant biases. Science is already politicized enough. Besides, in order for my assessment of costs to be adequate, I've got to get the science approximately correct first – wishing and hoping don't suffice (as Potti and Nevins discovered in Excursion 1).

We can agree with Ziliak and McCloskey if all they mean is that in deciding if a treatment, say hormone replacement therapy (HRT), is right for me, then a report on how it improves skin elasticity ignoring, say, the increase in cardiac risk, is likely irrelevant for my decision.

Some might eschew all this as naïve: scientists cannot help but hold on to beliefs based on private commitments, costs, and other motivations. We may readily agree. Oliver Lodge, our clairvoyant, had a keen interest in retaining a Newtonian ether to justify conversations with his son, Raymond, on "the other side" (Section 3.1). Doubtless Lodge, while accepting the interpretation of the deflection experiments, could never really bring himself to disbelieve in the ether. It might not even have been healthy, psychologically, for him to renounce his belief. Yet, the critical assessment of each of his purported ether explanations had nothing to do with this. Perhaps one could represent his personal assessment using a high prior in the ether, or a high cost to relinquishing belief in it. Yet everyone understands the difference between *adjudicating* disagreements on evidential grounds and producing a psychological conversion, or making it worthwhile financially, as when a politician's position "evolves" if the constituency demands it.

"Objective" (Default, Non-subjective) Bayesians

The desire for a Bayesian omelet while only breaking "objective" eggs gives rise to default Bayesianism or, if that sounds too stilted, default/non-subjective.[3] Jim Berger is one of the leaders:

> I feel that there are a host of practical and sociological reasons to use the label 'objective' for priors of model parameters that appropriately reflect a lack of subjective information ... [None of the other names] carries the simplicity or will carry the same weight outside of statistics as 'objective.'... we should start systematically accepting the 'objective Bayes' name before it is co-opted by others. (Berger 2006, p. 387)

[3] Aside: should an author stop to explain every joke, as some reviewers seem to think? I don't think so, but you can look up "omelet Savage 1961."

The holy grail of truly "uninformative" priors has been abandoned – what is uninformative under one parameterization can be informative for another. (For example, "if θ is uniform e^{θ} has an improper exponential distribution, which is far from flat"; Cox 2006a, p. 73.) Moreover, there are competing systems for ensuring the data are weighed more heavily than the priors. As we will see, so-called "equipoise" assignments may be highly biased. For the error statistician, as long as an account is restricted to priors and likelihoods, it still leaves out the essential ingredient for objectivity: the sampling distribution, the basis for error probabilities and severity assessments. Classical Bayesians, both subjective and default, reject this appeal to "frequentist objectivity" as solely rooted in claims about long-run performance. Failure to craft a justification in terms of probativeness means that there's uncharted territory, waiting to be developed. Fortunately I happen to have my own maps, rudimentary perhaps, but enough for our excavations.

Beyond Persuasion and Coercion

The true blue subjectivists regard the call to free Bayesianism from beliefs as a cop-out. As they see it, statisticians ought to take responsibility for their personal assessments.

> To admit that my model is personal means that I must persuade you of the reasonableness of my assumptions in order to convince you . . . To claim objectivity is to try to coerce you into consenting, without requiring me to justify the basis for the assumptions. (Kadane 2006, p. 434)

The choice should not be persuasion or coercion. Perhaps the persuasion ideal served at a time when a small group of knowledgeable Bayesians could be counted on to rigorously critique each other's outputs. Now we have massive data sets and powerful data-dredging tools. What about the allegation of coercion? I guess being told it's an Objective prior (with a capital O) can sound coercive. Yet anyone who has met Jim Berger will be inclined to agree that the line between persuasion and coercion is quite thin. His assurances that we're controlling error probabilities (even if he's slipped into error probability$_2$) can feel more seductive than coercive (Excursion 3).

Wash-out Theorems

If your prior beliefs are not too extreme, and if model assumptions hold, then if you continually observe data on H and update by Bayes' Theorem, in some long run the posteriors will converge – assuming your beliefs about the likelihoods providing a random sample are correct. It isn't just that these wash-out theorems have limited guarantees or that they depend on agents assigning non-zero priors to the same set of hypotheses, or that even with non-extreme

prior probabilities, and any body of evidence, two scientists can have posteriors that differ by arbitrary amounts (Kyburg 1992, p. 146); it's that appeals to consilience of beliefs in an asymptotic long run have little relation to the critical appraisal that we demand regarding the case at hand. The error statistician, and the rest of us, can and will raise criticisms of bad evidence, no test (BENT) regarding today's study. Ironically, the Bayesians appeal to a long run of repeated applications of Bayes' Theorem to argue that their priors would wash out eventually. Look who is appealing to long runs! Fisher's response to the possibility of priors washing out is that far from showing their innocuousness, "we may well ask what the expression is doing in our reasoning at all, and whether, if it were altogether omitted, we could not without its aid draw whatever inferences may, with validity, be inferred from the data" (Fisher 1934b, p. 287).

Take Account of "Subjectivities"

This is radically ambiguous! A well-regarded position about objectivity in science is that it is best promoted by excluding personal opinions, biases, preferences, and interests; if you can't exclude these, you ought at least to *take account of them*. How should you do this? It seems obvious it should be done in a way that *excludes biasing influences from claims as to what the data have shown*. Or if not, their influence should be made explicit in a report of findings. There's a very different way of "taking account of" them: To wit: view them as beliefs in the claim of inquiry, quantify them probabilistically, and blend them into the data. If they are to be excluded, they can't at the same time be blended; one can't have it both ways. Consider some exhibits regarding taking account of biases.

Exhibit (i): Prior Probabilities Let Us Be Explicit about Bias. There's a constellation of positions along these lines, but let's consider Nate Silver, the well-known pollster and data analyst. I was sitting in the audience when he gave the invited president's address for the American Statistical Association in 2013. He told us the reason he favored the Bayesian philosophy of statistics is that people – journalists in particular – should be explicit about the biases, prior conceptions, wishes, and goals that invariably enter into the collection and interpretation of data.

How would this work, say in *FiveThirtyEight*, the online statistically based news source of which Silver is editor-in-chief? Perhaps it would go like this: if a journalist is writing on, say, GM foods, she should declare at the outset she believes their risks are exaggerated (or the other way around). Then the reader can understand that her selection and interpretation of facts was through the lens of the "GM is safe" theory. Isn't this tantamount to saying she's unable to

evaluate the data impartially – belying the goal of news based on "hard numbers"? Perhaps to some degree this is true. However, if people are inclined to see the world using tunnel vision, what's the evidence they'd be able or willing to be explicit about their biases? Imagine for the moment they would. Suppose further that prior probabilities are to be understood as expressing these biases – say the journalist's prior probability in GM risks is low.

Now if the prior was kept separate, readers could see if the data alone point to increased GM risks. If so, they reveal how the journalist's priors biased the results. But if only the posterior probability was reported, they cannot. Even reporting the priors may not help if it's complicated, which, to an untutored reader, they always are. Further, how is the reader to even trust the likelihoods? Even if they could be, why would the reader want the journalist to blend her priors – described by Silver as capturing biases – into the data? It would seem to be just the opposite. Someone might say they checked the insensitivity of an inference over a range of priors. That can work in some cases, but remember they selected the priors to look at. To you and me, these points seem to go without saying, but in today's environment, it's worth saying them.[4]

Exhibit (ii): Prior Probabilities Allow Combining Background Information with Data. In a published, informal spoken exchange between Cox and me, the question of background information arose.

COX: Fisher's resolution of this issue in the context of the design of experiments was essentially that in designing an experiment you do have all sorts of prior information, and you use that to set up a good experimental design. Then when you come to analyze it, you do not use the prior information. In fact you have very clever ways of making sure that your analysis is valid even if the prior information is totally wrong. (Cox and Mayo 2011, p. 104–5)

MAYO: But they should use existing knowledge.

COX: Knowledge yes . . . It's not evidence that should be used if let's say a group of surgeons claim we are very, very strongly convinced, maybe to probability 0.99, that this surgical procedure works and is good for patients, without inquiring where the 0.99 came from. It's a very dangerous line of argument. But not unknown. (ibid., p. 107)

Elsewhere, Cox remarks (2006a, p. 200):

Expert opinion that is not reasonably firmly evidence-based may be forcibly expressed but is in fact fragile. The frequentist approach does not ignore such evidence but separates it from the immediate analysis of the specific data under consideration.

[4] Silver recognizes that groupthink created an echo chamber during the 2016 election in the USA.

Admittedly, frequentists haven't been clear enough as to the informal uses of background knowledge, especially at the stage of "auditing." They leave themselves open to the kind of challenge Andrew Gelman (2012) puts to Cox, in reference to Cox and Mayo (2011).

Surely, Gelman argues, there are cases where the background knowledge is so strong that it should be used in the given inference.

> Where did Fisher's principle go wrong here? The answer is simple – and I think Cox would agree with me here. We're in a setting where the prior information is much stronger than the data. . . . it is essential to use prior information (even if not in any formal Bayesian way) to interpret the data and generalize from sample to population. (Gelman 2012, p. 53)

Now, in the same short paper, Gelman, who identifies as Bayesian, declares: "Bayesians Want Everybody Else to be Non-Bayesian."

> Bayesian inference proceeds by taking the likelihoods from different data sources and then combining them with a prior (or, more generally, a hierarchical model). The likelihood is key. . . . No funny stuff, no posterior distributions, just the likelihood. . . I don't want everybody coming to me with their posterior distribution – I'd just have to divide away their prior distributions before getting to my own analysis. (ibid., p. 54)

No funny stuff, no posterior distributions, says Gelman. Thus, he too is recommending the priors and likelihoods be kept separate, at least for this purpose (scrutinizing an inquiry using background).

So is he agreeing or disagreeing with Cox? Perhaps Gelman is saying: don't combine the prior with the likelihood, but allow well-corroborated background to be used as grounds for scrutinizing, or, in my terms, conducting an "audit" of, the statistical inference. A statistical inference fails an audit if either the statistical assumptions aren't adequately met, or the error probabilities are invalidated by biasing selection effects. In that case there's no real disagreement with Cox's use of background. Still, there is something behind Gelman's lament that deserves to be made explicit. There's no reason for the frequentist to restrict background knowledge to pre-data experimental planning and test specification. We showed how the background gives the context for a FIRST interpretation in Section 3.3. Audits also employ background, and may likely be performed by a different party than those who designed and conducted the study. This would not be a Bayesian updating to a posterior probability, but would use any well-corroborated background knowledge in auditing. A background repertoire of the slings and arrows known to

threaten the type of inquiry may show a statistical inference fails an audit, or ground suspicion that it would fail an audit.

Exhibit (iii): Use Knowledge of a Repertoire of Mistakes. The situation is analogous, though not identical, when background knowledge shows a hypothesized effect to have been falsified: since the effect doesn't exist, any claim to have found it is due to some flaw; unless there was a special interest in pinpointing it, that would suffice. This is simple deductive reasoning. It's fair to say that experimental ESP was falsified some time in the 1980s, even though one can't point to a single bright line event. You might instead call it a "degenerating program" (to use Lakatos' term): anomalies regularly must be explained away by ad hoc means. In each case, Perci Diaconis (1978), statistician and magician, explains that "controls often are so loose that no valid statistical analysis is possible. Some common problems are multiple end points, subject cheating, and unconscious sensory cueing" (p. 131). There may be a real effect, but it's not ESP. It may be that Geller bent the spoon when you weren't looking, or that flaws entered in collecting, selecting, and reporting data. A severe tester would infer that experimental ESP doesn't exist, that the purported reality of the effect had been falsified on these grounds.

Strictly speaking, even falsifications may be regarded as provisional, and the case reopened. Human abilities could evolve. However, anyone taking up an effect that has been manifested only with highly questionable research practices or insevere tests, must, at the very least, show they have avoided the well-known tricks in the suitcase of mistakes that a researcher in the field should be carrying. If they do not, or worse, openly flout requirements to avoid biasing selection effects, then they haven't given a little bit of evidence – as combining prior and likelihood could allow – but rather an inference that's BENT. A final exhibit:

Exhibit (iv): Objectivity in Epistemology. Kent Staley is a philosopher of science who has developed the severity account based on error statistics (he calls it the ES account), linking it to more traditional distinctions in epistemology, notably between "internalist" and "externalist" accounts. In a paper with Aaron Cobb:

> . . . there seems to be a resemblance between *ES* and a paradigmatically externalist account of justification in epistemology. Just as Alvin Goldman's reliabilist theory makes justification rest on the tendency of a belief-forming process to produce true rather than false beliefs (Goldman 1986, 1999), *ES* links the justification of an inference to its having resulted from a testing procedure with low error probabilities (Woodward 2000). (Staley and Cobb 2011, p. 482)

The last sentence would need to read "low error probabilities relevant for satisfying severity," since low error probabilities won't suffice for a good test. My problem with the general epistemological project of giving necessary and sufficient conditions for knowledge or justified belief or the like is that it does not cash out terms such as "reliability" by alluding to actual methods. The project is one of definition. That doesn't mean it's not of interest to try and link to the more traditional epistemological project to see where it leads. In so doing, Staley and Cobb are right to note that the error-statistician will not hold a strictly externalist view of justification. The trouble with "externalism" is that it makes it appear that a claim (or "belief" as many prefer), is justified so long as a severity relationship SEV holds between data, hypotheses, and a test. It needn't be able to be shown or known. The internalist view, like the appeal to inner coherence in subjective Bayesianism, has a problem in showing how internally justified claims link up to truth. The analytical internal/external distinction isn't especially clear, but from the perspective of that project, Staley and Cobb are right to view ES as a "hybrid" view. In the ES view, the reliability of a method is independent of what anybody knows, but the knower or group of knowers must be able to respond to skeptical challenges such as: you're overlooking flaws, you haven't taken precautions to block errors and so on. They must display the ability to put to rest reasonable skeptical challenges. (Not just any skeptical doubts count, as discussed in solving induction in Section 2.7.) This is an integral part of being an adequate scientific researcher in a domain. (We can sidestep the worry epistemologists might voice that this precludes toddlers from having knowledge; even toddlers can non-verbally display their know-how.) Without showing a claim has been well probed, it has not been well corroborated. Warranting purported severity claims is the task of auditing.

There are interesting attempts to locate objectivity in science in terms of the diversity and clout of the members of the social groups doing the assessing (Longino 2002). Having the stipulated characteristics might even correlate with producing good assessments, but it seems to get the order wrong (Miller 2008). It's necessary to first identify the appropriate requirements for objective criticism. What matters are methods whose statistical properties may be shown in relation to probes on real experiments and data.

Souvenir P: Transparency and Informativeness

There are those who would replace objectivity with the fashionable term "transparency." Being transparent about what was done and how one got

from the raw data to the statistical inferences certainly promotes objectivity, provided I can use that information to critically appraise the inference. For example, being told about stopping rules, cherry picking, altered endpoints, and changed variables is useful in auditing your error probabilities. Simmons, Nelson, and Simonsohn (2012) beg researchers to "just say it," if you didn't p-hack or commit other QRPs. They offer a "21 word solution" that researchers can add to a Methods section: "We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study (p. 4)." If your account doesn't use error probabilities, however, it's unclear how to use reports of what would alter error probabilities.

You can't make your inference objective merely announcing your choices and your reasons; there needs to be a system in place to critically evaluate that information. It should not be assumed the scientist is automatically to be trusted. Leading experts might arrive at rival statistical inferences, each being transparent as to their choices of a host of priors and models. What then? It's likely to descend into a battle of the experts. Salesmanship, popularity, and persuasiveness are already too much a part of what passes for knowledge. On the other hand, if well-understood techniques are provided for critical appraisal of the elements of the statistical inference, then transparency could have real force.

One last thing. Viewing statistical inference as severe testing doesn't mean our sole goal is severity. "Shun error" is not a terribly interesting rule to follow. To merely state tautologies is to state objectively true claims, but they are vacuous. We are after the dual aims of severity and informativeness. Recalling Popper, we're interested in "improbable" claims – claims with high information content that can be subjected to more stringent tests, rather than low content claims. Fisher had said that in testing causal claims you should "make [your] theories elaborate by which he meant . . . [draw out] implications" for many different phenomena, increasing the chance of locating any flaw (Mayo and Cox 2006, p. 264). As I see it, the goals of stringent testing and informative theories are mutually reinforcing. Let me explain.

To attain stringent tests, we seek strong arguments from coincidence, and "corroborative tendrils" in order to triangulate results. In so doing, we seek to make our theories more general as Fisher said. A more general claim not only has more content, opening itself up to more chances of failing, it enables cross-checks to ensure that a mistake not caught in one place is likely to ramify somewhere else. A hypothesis H^* with greater depth or scope than another H may be said to be at a "higher level" than H in my horizontal "hierarchy" (Figure 2.1). For instance, the full GTR is at a higher level than the individual

hypothesis about light deflection; and current theories about prion diseases are at a higher level than Prusiner's initial hypotheses limited to kuru. If a higher level theory $H^{\star}$ is subjected to tests with good capacity (high probability) of finding errors, it would be necessary to check and rule out more diverse phenomena than the more limited lower level hypothesis H. Were $H^{\star}$ to nevertheless pass tests, then it does so with higher severity than does H.