3 LEARNING FROM ERROR: HOW EXPERIMENT GETS A LIFE (OF ITS OWN)

Deborah G. Mayo

'Experiment Lives a Life of its Own'

This familiar slogan is often thought to capture the epistemic importance and power attributed to experiment. But decades after Hacking and others popularized it in philosophical circles, we are still in need of building a full-bodied philosophy of experiment.

- How should we understand this slogan?
- Where does the separate life of experiment take place?
- How does it manage to get a life of its own?
- Why should it want its own life?

Beginning with the first question, we may consider three interrelated glosses on the 'own life' slogan.

Experimental Aims: Apart from Theory Appraisal

The first sense in which experiment may be said to have its own life concerns experimental aims: to find things out quite apart from testing or appraising any theory. The goals are the local ones of obtaining, modelling and learning from experimental data: checking instruments, ruling out extraneous factors, getting accuracy estimates, distinguishing real effect from artefact, signal from noise. Experiments are often directed at taking up the challenge of designing better experiments: how can we learn more, and do it faster? How can we more cleverly circumvent flaws and limitations?

To begin with, researchers wish to explore whether there is even something worth investigating. There may be no theory in place to flesh out, much less to test. The very domain in which any eventual theory might live may be unclear. Even when the goal is to fill out or test a theory, intermediate experimental inferences are needed to bridge the data-theory gaps. Experimental phenomena may concern effects that the theory does not even talk about — an important part of
their power. Theories, even where we have them, do not tell us how to test them. Nor will it do to imagine that theories are tested by joining various auxiliary hypotheses, hypothetico-deductively. Even in the rare cases where scientists can do that, it would not be a profitable way to go.

The aims of experiment – both in day-to-day practice, and in the startling discoveries on the cutting edge – are most aptly characterized by the simple desire to find things out.

**Stability and Stubbornness: Experimental Knowledge Remains**

The second gloss on the slogan concerns the continuity and growth of experimental knowledge. Despite changes in theory, even in the face of the need to reinterpret the significance of experimental inferences, claims that pass a stringent ‘test of experiment’ are generally stable. Even if the goal of the experiment is to estimate the parameters of a theory, the crux of good experiments is that the inferred estimates may be detached reliably. They are (as experimental physicists say) ‘clean estimates’ not sullied by unknowns. The deflection of light parameter in experimental general relativity, for example, is a clean effect that any adequate theory of gravity will have to accommodate.

Experimental knowledge grows, as do the tools for acquiring it, whether instrumentation, manipulation, computation, self-correction and so on. Continuity at the level of experimental knowledge points to a crucial kind of progress that is overlooked when measures of growth are sought in terms of large-scale theory change, in updating probability assignments to theories, or by means of other favoured macro-methodology schemes.

**Independent Warrant for Inference**

This leads to the deepest point about the ‘own life’ slogan: namely, the independent justification of experimental data and inference. Independent of what? The answer generally given is: independent of theory or at least ‘high level’ theory, inviting the criticism that theory always enters. But that mistakes what the ‘own life’ achievement is all about. What really matters is attaining freedom from whatever could be a threat to what the researchers are trying to find out. The interesting thesis is that experimental evidence and inference need not be theory-laden in any way that invalidates their various roles in grounding experimental arguments. They need not be dragged down by whatever is thus far unknown. While granting that experimental data are not given unproblematically, the position is that an experimental inference may be vouchsafed apart from threats of error. A particular experimental inference may be in error, but so long as errors are sufficiently understood and controlled we may discover, and avoid being misled by, them.

Some philosophers of experiment stress the independent grounding afforded by knowledge of instruments; others stress the weight of certain experimental activities such as manipulation. But neither is necessary or sufficient for experimental learning. Knowledge of instruments and astute manipulation may be important in obtaining experimental knowledge, but any means of reliable experimental learning may do as well or better, including computation, simulation, and statistical methods and modelling.

The term ‘reliable’ is notoriously ambiguous. I will put it to one side for now except to note that it does not suffice that the experiment will get it right in the long run, with high probability, asymptotically, or the like. A reliable experiment has to be capable of controlling misinterpretations in the case at hand, or at least within the time of a typical inquiry, research effort or report. The goal is not error avoidance but error control, which may be had by deliberately capitalizing on ways we know we can be wrong.

**Canonical Errors: Terminology**

In speaking of ‘errors’, I am not referring merely to observational errors, systematic or unsystematic, but rather mistaken understandings of any aspect of a phenomenon. Some may see this as a nonstandard use of ‘error’. The ways that a claim or hypothesis H may be ‘in error’ include erroneous claims about underlying causes and mistaken understandings of any testable aspect of a phenomenon of interest. I am not drawing distinctions between experimental and theoretical, as some might. I am prepared to call a context ‘experimental’ (whether literal manipulation is present or not) insofar as error-probing capacities can be controlled and assessed.

Co-opting a term from (frequentist) statistics, where the probabilities of methods for discerning errors are called error probabilities, I refer to an error-statistical approach (though the account is certainly not limited to formal statistical experiments).

While there are myriad types of mistakes in inference, I propose that there are a handful of error types and strategies for checking, avoiding and learning from them. I term these ‘canonical’ errors:

- mistaking chance effects or spurious correlations for genuine correlations or regularities
- mistakes about a quantity or value of a parameter
- mistakes about a causal factor
- mistakes about the assumptions of the data (for the experimental inference)
- mistakes in linking experimental inferences and subsequent claims or theories.
While these are not exclusive (for example, checking assumptions and causal factors may take the form of testing for parameter values couched in models), each seems to correspond to distinct types of arguments and standards of evidence. Even so, I want to emphasize that these are just ways of categorizing strategies for far more context-dependent queries. There is a corresponding localization of what one is required to control, as well as what one is entitled to infer.

The interest is in capturing, quite generally, erroneous inferences: erroneous interpretations of data or erroneous understandings of phenomena and whatever can hinder the discovery of them. In one sense the goal is deeper and more interesting than getting a true hypothesis or theory. Even if, say, experimental relativists knew in 1930 that general relativity was true, they could not be said to have 'correctly understood' (relativistic) gravity. That demanded probing how gravity behaves in specially designed experiments.

Paying deliberate attention to errors, I claim, is at the heart of getting correct inferences and warranted interpretations of data. (So, clearly, the topic of this forum is near and dear to me.) Even granting that all models are wrong, and all theories strictly false, we are not prevented from getting a correct understanding of experimental stabilities and effects: they have their own life.

When Is It Bad for Data to Depend on Theory?

The most serious problem with theory-dependent inferences arises when the very hypothesis or theory under test is implicitly assumed. If it is predetermined that the interpretation of data is constrained to be in accord with a hypothesis \( H \), whether or not \( H \) is correct, then the fact that data accord with \( H \) scarcely counts in its favour. This is to use a method that has no way of teaching us where \( H \) may be wrong. Enterprises that regularly proceed in this way would be considered pseudoscientific: data are not being taken seriously in the least. Even if the data set \( x \) accords with or 'fits' \( H \), the procedure has maximal error probability: \( H \) 'passes' a test with minimal stringency or severity:

If data \( x \) is generated by a method or procedure with little or no capability of finding (uncovering, admitting) the falsity of \( H \), even if \( H \) is false, then \( x \) is poor evidence for \( H \).

However, it may not be at all obvious, at least in interesting cases, that a method is guilty of such prejudgement. Whether data-dependent inferences (data mining, non-novel data) lead to high error probabilities needs to be scrutinized case by case.

For instance, many think that the fact that a hypothesis \( H \) was constructed to fit or accommodate the data, prevents the data also from counting as a good test of, or warranted evidence for, \( H \). This 'no double-counting' requirement cap-

tures a general type of prohibition against data mining, hunting for significance, tuning on the signal and ad hoc hypotheses, in favour of requiring predesignated hypotheses and novel predictions. Whether (and when) inferences should be discounted, if not disallowed, when double counting has occurred has long been the source of disagreement and debate in philosophical and statistical literature. It is well known that if one is allowed to search through several factors and selectively report only those that show (apparently) impressive correlations, then there is a high probability of erroneously inferring that a correlation is real. However, it is equally clear that there are procedures for using data both to identify and test hypotheses that would only rarely output false hypotheses: the use of a DNA match to identify a criminal, radiointerferometry data to estimate the deflection of light, and entirely homely examples such as using a scale to measure my weight. Here, although the inferences (about the criminal, the deflection effect, my weight) were constructed to fit the data, they were deliberately constrained to reflect what is correct, at least approximately.

The secret to distinguishing cases, I argue, is staying clear about what error one needs to worry about, and whether the severity, stringency or probativeness of the test to avoid the error is compromised. What matters is not whether \( H \) was deliberately constructed to accommodate data \( x \), but how well the data, together with background information, rules out ways in which an inference to \( H \) can be in error.

Severity Plus Informativeness: Experimental Kuru

Merely avoiding error is scarcely sufficient for finding things out. True, before inferring \( H \) we want the experiment to have had a fairly high capability of unearthing flaws in \( H \) (stringency, severity), but aside from stringency we want informativeness. A key experimental aim is to carry out inquiries that will enable us to find something out, to extend our knowledge. The test or inquiry should have a fairly good chance of teaching us something. Or else we could stick to uttering tautologies and making unrisky claims. We are prepared to risk error because we want to learn. We deliberately construct experimental phenomena in order to learn about naturally occurring phenomena, but in such a way that error is controlled.

Consider the example of a brain disorder known as kuru (derived from 'to shake'), which ravaged the Fore people of New Guinea during the 1950s and 1960s. In its clinical stage, which lasts an average of twelve months, its sufferers go from having difficulty walking, to laughing uncontrollably, to being unable to swallow, to dying. Kuru and (what we now know to be) related diseases like BSE (mad cow disease), Creutzfeldt–Jakob disease and scrapie are 'spongiform' diseases—they cause the brain to appear spongy. (They are also called TSEs: transmissible spongiform encephalopathies.)
D. Carleton Gajdusek probed the disease experimentally without a clue about what a satisfactory theory of kuru might be. He asked: what causes kuru? Is it transmitted through genetics? Infection? How can it be controlled or eradicated? No theory is needed to begin to ask how hypothetical answers to these questions might be wrong.

There was considerable data in the 1950s that kuru clustered within families, in particular among women and their children, or elderly parents. Might it then be genetic, like Alzheimer's disease, which produces similar plaques in the brain? But that was soon ruled out as erroneous, if only because it would have killed off the tribe. Researchers began to suspect that kuru was transmitted through mortuary cannibalism to maternal kin, a practice that both honoured the dead and let women in on the limited meat supply. Ending the cannibalism all but eradicated what had been an epidemic disease. This alone did not show the correctness of the mortuary cannibalism hypothesis. (Many of the tribe remained convinced it was witchcraft.)

In order to start experimenting, they had to dream up a method that would allow delimiting and probing possible causes through deliberately triggered experimental versions of kuru: could kuru be transmitted to animals if their brains were inoculated with infected tissue from kuru victims?

The similarities between kuru and scrapie in sheep led Gajdusek to begin experiments that would test whether kuru could be transmitted to animals. Many of the attempts to transmit kuru failed, but given the difficulty of the process, the length of transmission and much else, the failures showed only the inadequacy of the experiments: they were not evidence of the falsity of the kuru hypothesis. Nor were they uninformative; they were the basis for improved experiments. Years of lengthy experiments would show that kuru could be experimentally transmitted to chimpanzees, monkeys and other animals if they were injected with specially prepared infected brain extracts (from kuru victims). They studied experimental kuru to explore actual kuru.

Experimentalists did not pretend to have gathered anything more than evidence about kuru's transmission; they called it a 'slow virus' (given the lengthy incubation), understood as little more than whatever it was they were transmitting. A report of errors not yet probed was an important part of the research report. They had at most a family of hypotheses about which to inquire. But researchers were excited by the fact that posing questions about deliberately triggered experimental phenomena had opened up platforms for engaging with a variety of TSEs. What was being learned lived a life of its own apart from theories.

No one expected that what had been learned about kuru would have revolutionary implications for a novel type of infectious particle (with no nucleic acid) that Stanley Prusiner would term a prion (proteinaceous infectious particle).

Arguing From Error and Revolutionary Science

That experimental knowledge may get a life of its own is clearly relevant to avoiding a problem that gives philosophers of science so much trouble: theory-laden data. Kuhn contrasted revolutionary science with so-called normal science, wherein researchers flesh out theories and solve articulated problems or 'puzzles' posed by the aims and methods of the large-scale theory or 'paradigm.' While few philosophers nowadays accept the full Kuhnian (or even the Lakatosian) picture, it seems they refuse to shake off one or more of its stumbling blocks. Many think that any anomaly can be avoided by changing enough of the background; while logically true, such a tactic is a poor way of learning.

A well-known Kuhnian position is that while paradigms provide the researcher with tools for conducting science within a paradigm, they supply no tools for breaking out of that very paradigm: 'Paradigms are not corrigible by normal science at all!' Not so. Local experimental results may show unavoidable anomalies that thereby overthrow hard cores with the same methods at hand. In one sense, I am simply taking seriously Kuhn's idea that 'severity of test-criteria is just one side of the coin whose other face is a puzzle-solving (i.e. a normal science) tradition.' Where we differ, or seem to, is that I deny that there is a break when it comes to correcting theories. Genuine experimental knowledge is stable, and any future theory of the phenomenon, to be adequate, must accommodate it.

Overthrowing the Central Dogma of Biology

Reproducible data began to accumulate, indicating that whatever was causing kuru could not be eradicated with techniques known to kill viruses and bacteria; furthermore, victims were not showing the presence of antibodies that would be produced by infectious elements that possessed nucleic acid. At the same time, researchers observed that infectivity (of both kuru and scrapie) was weakened by factors known to modify proteins.

As the experimentalists described in detail, prions, whatever they may be, resist inactivation by UV irradiation, boiling, hospital disinfectants, hydrogen peroxide, iodophors, peracetic acid, chaotropes and much else. So if it were a mistake to construe kuru as having no nucleic acid, then at least one of these known agents would have eradicated it. A general pattern of argument emerges.

Experimental Argument from Error

We argue that H is a correct construal of data x when the procedure would have unearthed or signalled the misinterpretation, but instead regularly produces results in accordance with H. This is an example of what I call an argument from
error.11 (Note: data x would generally be a vector of outcomes possibly from several sub-experiments.)

They did not know what this non-virus was, and Prusiner said as much ('the transmission of experimental Kuru became well established, but its mode of action remained puzzling'). The genius of experiments is precisely in their allowing us to ask a single question at a time: Nucleic acid or not? Protein or not? Whatever it was, it appeared to be an infectious substance that was neither a virus nor bacteria, at odds with 'the central dogma' of biology. There was a genuine experimental phenomenon that would not go away, as well as a platform for launching ever-deeper probes of error.

It's All in the Planning (and Design) of Experiments

Experiment living 'its own life', circumventing obstacles to finding things out, is a consequence of something that has received too little attention in philosophies of experiment: experimental design and planning. I include under this rubric any aspect of deliberate planning for the collection, 'treatment', modelling and analysing of data. Experiments are distinguished from passive observation precisely because of the role of deliberate design in delimiting and controlling factors that would otherwise interfere with learning about effects of interest. Even with 'fortuitous' observations, researchers may try to mimic what experiment offers.

This brings up the question of what kinds of examples philosophers of experiment might fruitfully consider in constructing an experimental account. If perfect controls were attainable then we would not be in a very illuminating domain of learning, so I am interested in more challenging kinds of cases. For the same reason, we can learn the most by looking beyond cases where the primary goals are safely limited to a particular data-generating mechanism. It is precisely for this reason that I am often led to considering cases where the best one can do is to model experiments statistically. Statistically framed inquiries, by being explicit about the fact that the primary question is being 'embedded' into a statistical data-generating mechanism (DGM), emphasizes the need to connect back 'up' from what is learned about the DGM to the substantive or primary question of interest.

Homes for Experimental Life

Experiments have lives of their own, but it should be a real life, not a life in the street, with its own parameters, models, theories. An account that begins with given statements of evidence and hypotheses will not be relevant to the actual practice of experimental science. This recognition has been one of the most important ways that experimental philosophers have revealed the shortcomings of accounts that seek evidential-relation 'logics' assumed to hold between any given statements of data and hypotheses. That an adequate account of experi-

What about the work that goes into designing (or specifying) hypotheses to test or infer? Much less has been said about this and it is a central gap which I encourage philosophers of experiment to fill. Leaving issues of discovery and specification vague is one of the reasons for the lack of clarity on the issue of when data-dependent specifications and selection effects matter. It has also led philosophers of induction and evidence to overlook the crucial value of very local, piecemeal hypotheses 'on the way' towards arriving at substantive claims and models.

I want to highlight a triad of components in planning, running and interpreting experiments in practice, including:

- questions, hypotheses
- data collection
- data analysis and interpretation.

There may be only a vague question or a loose family of speculative hypotheses (pre-data), and these typically will be very different from hypotheses or inferences arrived at (post-data). To capture a single unit of experimentation (perhaps as big as a given inquiry or published paper) with plenty of roomy niches for these components (see Figure 3.1):
However, the three components are actually intimately connected, and the planning concerns all three at once. What is inferred at one stage may well be data for the next. Moreover, the deliberate use of raw data, differently modelled, is crucial to the problem of testing assumptions for a given experiment. The details by which experiments can self-correct their own assumptions should be a central component of any adequate account of experimental learning.

C. S. Peirce: Quasi-Experiments

It is perhaps unsurprising to find support for this conception in the work of Charles Peirce, who promoted the idea that the key to inductive inference was the ability to self-correct and learn from error. Peirce, himself an experimental scientist, coined the term 'quasi-experimentation' to cover the entire process of searching, generating and analysing the data, and using them to test a hypothesis and 'this whole proceeding I term induction'.

Further, for Peirce, the 'true and worthy' task of logic was to 'tell you how to proceed to form a plan of experimentation'. Were we to take this Peircean idea seriously, we would get beyond sterile a priori logics of evidence to something much more interesting and relevant.

Appraising/Ensuring Error-Probing Capacities

The key question directing experimental planning is how to split off questions that can be probed in terms of experimental data-generating procedures that will afford adequate control of interfering factors. The main considerations are these:

1. to ensure (pre-data) that experimental analysis of the data will be possible so that something is very likely to be learned, even if the experiment is botched.
2. to design 'custom-made' effects that might not even exist in nature in order to make the data talk to us (about aspects of their origins).

Design specifications alter the error-probing capacities of experiments, but far from sullying the results, they are the key to an objective interpretation of results. In statistics we use characteristics of the experimental test to determine the probability that a discrepancy from a test hypothesis will be inferred erroneously (type 1 error), and that discrepancies will be erroneously overlooked (type 2 error). The analogous critique is conducted informally in scrutinizing experiments.

Experimental Accounts Should be Objective While Empirical

Even with this more complex picture of experimental ingredients, we need not forgo the philosopher's yen to identify overarching patterns of argument. The argument patterns follow variations on the basic structure of learning from error. The nitty-gritty details involved in cashing out such arguments — organized by a handful of error types — call for empirical experimental knowledge. Yet the account itself is normative.

Given especially the consciousness raising afforded by the focus on experiment, it is odd still to hear some philosophers maintain that if we relinquish the idea of a logical relationship between given evidence and hypotheses (evidential-relation logics), then we also give up on objectivity. Achinstein,15 for instance, declares that philosophical accounts of evidence are irrelevant to scientists because they are a priori while scientists evaluate evidence empirically. I think this is a serious mistake.

Popper recognized that there was a tension between our intuitions about stringent tests that avoid being ad hoc, and the aim of constructing formal logics of evidence, but fell into the same false dilemma. While Popper famously spoke of corroboration as evidence that passed sincere attempts to falsify, he was at a loss to make operational 'sincere attempts to falsify', although he clearly did not intend it to be a matter of psychology. Merely 'trying' to falsify, even if we could measure effort, would be completely beside the point: one must deliver the goods. Clearly, the Popperian idea that claims should be accepted only after passing 'severe attempts to falsify' is in the error statistical spirit; but Popper never had an account of evidence that could do justice to this insight.16 His followers, the critical rationalists, have not gone further except to note that we need to go beyond formal logics to recognize what Lakatos and Musgrave call the 'historical' nature of evidence.17

I heartily agree that we need to take into account how hypotheses are constituted and selected for testing. But too often what one counts as evidence is thought to be subjective, relative, historical or psychological. There are equivocations here of which we must steer clear: different stages of inquiry, just like different regulatory bodies (e.g. OSHA versus the EPA in the USA), may use different standards, but given the chosen standard, whether or not it is met need not itself be subjective, historical or psychological. That there is a multitude of different, and context-dependent, ways to satisfy requirements of experimental evidence does not imply that judging whether evidence is good or poor is a relative matter.18

In order to appraise purported claims of evidence, without which the context is not properly experimental, one must scrutinize the (error-statistical) properties of experimental tools and methods, their capacity or incapacity to distinguish errors and discriminate effects. Combining tests strengthens and fortifies the needed analysis. This is an empirical task all right, but it can be free of debilitating assumptions, which is at the heart of the 'own life' achievement.
How Experiments Free Themselves

An experiment need not actually exclude interfering factors to be 'free' of them. It suffices to estimate (or simulate) their influence, 'knock out' or subtract out effects, or cleverly distinguish or disentangle the effects of different factors. This gives rise to a variety of strategies, notably:

Amplifying effects:

An excellent way of learning from error is to deliberately magnify it. A pattern may readily be gleaned from noisy data by introducing known standards and measuring deviations or residuals. By magnifying distortions, mere whispers can be made to speak volumes. Through simulations and statistical modelling we may find out what it would be like if a hypothesized construal of data were wrong, or if a given factor were operating.

Self-correcting and/or showing robustness:

It is often still said that an experimental inference is only as reliable as the sub-experiments involved; but with good experiments, garbage in need not be garbage out. The overall inference may be far more reliable than any individual sub-experiment. Experiments achieve lift-off. Violating background assumptions might render them less efficient, but the cornerstone of good experiments is to allow the inference still to stand (robustness). By contrast, if satisfying assumptions of data and experimental models are highly sensitive to minor or common errors in attainable data, the experimental inferences have a hard time getting free of them. Fortunately, techniques for discovering and even exploiting such obstructions can be deliberately introduced into shrewdly crafted experiments.

Appealing to repertoires of error:

Clearly, any particular experimental triad \( E_i \) uses and builds on a general background of information which I dub a repertoire of errors. This would include errors that have and have not been ruled out in the area of interest, as well as pitfalls and success stories in dealing with the type of error that may be of concern. An adequate philosophical account of experiment should include such repertoires; there is considerable latitude as to how this may be accomplished. Historical cases may illuminate instances where errors were made several times before they were canonized. New ideas for cross-checking could well emerge.\(^1\)

If we have put together a sufficiently potent arsenal for unearthing types of mistakes, we may construct a testing procedure that is highly capable of revealing the presence of a specific error, if it exists, but not otherwise. We may call this a \textit{severe error probe}. If no error is uncovered with a demonstrably severe error probe, there is evidence it is absent. By contrast, if no error is discerned, but the procedure was fairly likely to miss it, then we are not warranted in inferring it is absent, at least on the basis of this particular probe. (Finding no evidence of error is not evidence of no error.) Finding such insensitivity is itself highly informative: properly used, it offers one of the most effective ways to invent new hypotheses. In particular, it supplies ideas for hypotheses we would not have been able to discriminate with existing tests. We quite literally learn from error — more precisely, from recognizing the limited error discriminating abilities of a given test.

Experiments Can 'Use' but Still Not Depend on Theories

Some have complained that experimental philosophers 'were throwing the baby out with the bathwater' in their emphasis on low-level experimental hypotheses.\(^2\) We overlook, they say, how we invariably use background theories. What they overlook is how experiment may 'use' theories, or make use of theoretical knowledge, while not being threatened by them. They have their own life!

There is no problem in 'relaying' on background theories:

- to hypothetically draw their consequences for testing, or
- when they have passed severe tests of their own, or
- when the only aspects being relied on are known to hold sufficiently.

Finally, an experiment can free the interpretation from assumptions that might introduce error (theoretical or other); namely, state them in the sum-up. (This is akin to a conditional proof in logic.)

Some might find it unusual for a philosophical account of evidence to require explicit consideration of both what has and has not been adequately learned from a given experiment; but that is what I am proposing. It will not appeal to neatniks! But the complexity pays off in providing an account that does justice to the cleverness of experimental practice, and at the same time enables long-standing philosophical problems of evidence to be resolved.

Agnosticism on Scientific Realism

Underlying the 'throwing out the baby with the bathwater' remark, in some cases, is the assumption that anyone who does not set out to build an account of theory acceptance is an anti-realist and limited to 'empirical adequacy' or a mere 'heuristic' use of theories. Musgrave wonders why I am prepared to accept experimental claims but not theoretical ones; but as noted at the outset: I never made any such distinction.\(^3\) The models linking data to hypotheses and theories all contain a combination of observational and theoretical quantities and factors, as I see them. Since the same issues of warranting experimental knowledge arise for realists (of various stripes) and non-realists, the entire issue of scientific realism is one about which the error statistician may remain agnostic.
Experiment is Piecemeal: Threats of Alternative Hypotheses Squashed

Being able to implement these strategies leads to ‘getting small’, to specifying a question that will restrict or control erroneous interpretations of the kinds of data we are actually in a position to collect. This circumvents the familiar problem of ‘alternative hypothesis objections’.

A typical challenge is: how can you rule out ways $H$ can be false when there are always members in the ‘catchall hypothesis’ – including claims not even considered? To determine the likelihood of data under the catch-all hypothesis, as Wesley Salmon puts it, ‘we would have to predict the future course of the history of science’.22 The catch-all is indeed problematic for accounts that require listing all alternatives to a hypothesis $H$, or all the ways $H$ might be false. But experiment deliberately delimits and reduces the factors that could be responsible for observed effects. With effective experimental design and data generation, many of the logically possible explanations of results may be effectively rendered actually or virtually impossible. If a certain gene is successfully knocked out of a mouse, it is not responsible for the effect; if the researchers do not know who got a placebo, then preconceptions of effectiveness are removed or at least diminished. The goal of experimental design is to specify a question that will restrict or control erroneous interpretations of the kinds of data we are actually in a position to collect. We can exhaust the space with respect to that one question. Even if we err, we may arrange things so that there is a good chance it will be detected in subsequent checks or attempts to replicate.

Rather than trying to distinguish a hypothesis or theory from its rivals, experiment sets out to distinguish and rule out a specific erroneous interpretation of the data from this experiment. In so doing the experimenter is not restricted by initial questions posed; a post-data appraisal is needed. Given the data, experimenters ask: what, if anything, do these data enable ruling out?

One could rationally reconstruct experimental inquiry using models of large-scale theory change, of Bayesian updating, or of decision theory with specified losses – once an episode is neat and tidy. The ease of doing so, some think, is one of the weaknesses of such reconstructions. Like a paint-by-number algorithm for the Mona Lisa, they do not capture how the learning took place (or the painting was created). They are backward looking, not forward looking, and fail to do justice to actual experiments. In setting sail to find things out, pre-data specifications of an exhaustive set of rival substantive hypotheses are atypical; much less do we have, or want to interpret the data in light of, cost or benefit functions as a decision-theoretic construal requires. Even evidence-based policy (or other subsequent decisions) should rest on a valid experimental knowledge base.23 Certain low-level claims (about parameters, directions of effect, observed correlations) may appear overly simple if they are thought to be the final object of study, but for exhausting the space of a local error (e.g., $\mu < \mu'$ or $\mu > \mu'$) they are just the ticket!

Arguing from Coincidence

The powerful form of argument that experiments provide is often described as an ‘argument from coincidence’: there is no way that all of these well-known instruments and independent manipulations could consistently produce certain effects or concordant data, were they all artefacts of instruments. It is not merely that the concordant results are formally improbable. As Hacking notes, to suppose that they are all instrumental artefacts is akin to invoking a Cartesian demon of the instruments.24 This inference to a non-artefact (or ‘real effect’) is an instance of my general argument from error: ‘We argue that there is evidence that (the artefact’s) error is absent when a procedure that would have (with high probability) unearthed the (artefact’s) error fails to do so, but instead consistently produces results indicating its absence.’

This can also be put in terms of inferring a genuine, non-chance effect.25 We may deliberately create artefacts and discern how readily they are revealed. This teaches us about the error-probing capacities of our overall experiment, which combines many results. This is all part of the background repertoire of knowledge of errors.

Knowing that an effect is ‘real’, in the sense of non-artefactual, however, is one of the strongest kinds of experimental knowledge. Still it is important. It was the first error on my list. It may arise to check whether an instrument is working, or even to ascertain whether there is any real effect worth exploring. Experimenters want to know if they would be wasting their time trying to explain effects that could readily be accounted for by chance.

Arguing from Coincidence to What?

By leaving arguments from coincidence at a vague level, however, they are often appealed to as warranting much more than they actually allow. Avoiding the error that needs avoiding in order to infer that there is evidence of a real (not spurious) effect does not directly warrant hypothesized explanations of the effect. (That is why statistical significance is not substantive significance.)

Yet many say they do: would it not be preposterous coincidence if all these different experiments $E_1, E_2, E_3, \ldots, E_n$ yielded data $x$ in agreement with theory or hypothesis $H$, if $H$ were false? Would it be? That is what we would need to figure out. If it would, then the argument shows $H$ to be well or severely tested in my sense. If it would not, we should not be accepting all of $H$ on the basis of data $x$.

To suppose we should, or that scientists do, is to take a pattern of argument that works where we can exhaust, and rule out, all the possibilities and apply it in general. Instead I say we should put our epistemology of experiment at the level of experiment. Such local experiments offer standard or canonical ways of exhausting answers, e.g., nucleic acid or not? Far from seeking to infer all of $H$ before its time, the engine for experimental knowledge grows through understanding why and when we would not be warranting in doing so.
Relevant Variability Depends on the Error We Need to Rule Out

Most importantly, the varied experiments $E_1$, $E_2$, $E_3$, ..., $E_n$ must be shown to be relevantly varied! Errors should have ramifications in at least some of the other experimental trials. Whatever might threaten one experiment must not also be able to be responsible for an error in the others. If all witnesses have been bought off, a 'wide variety' of them may not yield anything more probative than one; if all the samples are contaminated, their agreement does not help.

If our experimental account is to be forward looking, as I urge, then the focus should be on how to move, in the constrained fashion that experiments allow, from what is known at a given stage to learning more. These may be baby steps, but at least they will be taken securely. Transmission of kuru was known to be real in the 1980s, say, but do the data warrant a 'protein-only' explanation? No. And experimenters could not even experimentally probe such a question before learning how to construct testable forms of the 'protein-only' hypotheses. So let us turn to some aspects of experimental learning about prions. I obviously cannot here relate this rich and decades-long episode; my goal is to uncover a few of the gems it offers for experimental learning from error.

An Experimental Platform for Understanding Prion Diseases

Key aspects of what enables experiments to live lives of their own are exemplified by the use of 'animal models' in probing the transmission of kuru and other TSEs (in this case mostly in mice and hamsters): the models supply a general platform on which to probe various aspects of a phenomenon of interest (here, prion transmission) so that what is learned remains regardless of subsequent reinterpretations or overarching theories still to emerge.

As the researchers got better at purifying prions, it became clear that the minimum molecular weight necessary for infectivity was so small that it excluded viruses and any other known infectious agent. Prions were found to contain a single protein dubbed PrP. To the researchers' surprise, PrP was found in normal animals - so it does not always cause disease. They dubbed the non-pathogenic, common form PrP-C; the pathogenic, or scrapie form, PrP-Sc. Then they deliberately designed experiments to discover if they 'had made a terrible mistake' (and prions had nothing to do with it). They had not. One way of learning how matters would have gone had they been wrong, however, was to create transgenic mice with PrP deliberately knocked out. No such knock-out mice were able to be infected with PrP-Sc.

Getting Good at Learning about Prion Transmission

But how is it that different patterns of infection were observed, with different incubation times and apparent species barriers? The answer seems to point to different 'strains' of prions, but strains are the sort of thing only viruses were thought to have.

The researchers did not know what they would find, and an adequate (forward-looking) account of experiment should be at home with this fact of experimental life. What they can and did do is design experiments that would give them a good chance of learning something about species barriers to infectivity.

Transgenic mice with a hamster PrP gene were created: when inoculated with mouse prions, they made more mouse prions; when they were inoculated with hamster prions, more hamster prions. They were onto something that might explain the species barrier. Mice are not normally infected with hamster prions, but hybrid mice, created with portions of both hamster and mouse protein sequences, they discovered, could be infected with either mouse or hamster proteins! By combining transgenic approaches and computer modelling methods, they were able to produce mice susceptible or resistant to prion disease in predictable ways. This was the basis for experimental learning.

Decades before they had begun to understand the mechanism behind the observed species barriers in infection, prion transmission between species was modelled as a stochastic process (based on scrapie in sheep). During the 1980s, experimental manipulation enabled them to predict and control transmission, so that 'it becomes a nonstochastic process'. They knew their understanding of transmission was growing when they could reduce the incubation period from 600 days to 90 with continual injections of pathogenic brain tissue. (Understanding different patterns of species transmission is obviously relevant to the question of BSE/mad cow disease in humans - when and why humans are infected, and how to detect it.)

For biologists, yeast is a great model organism: it has a growth cycle of only eighty minutes. But, as is typical, damping down one error heightens another. Only by connecting the transgenic and normal animal models, the yeast and the hamster, and so on, was it possible for one experiment to serve as a relevant check on the others, permitting the overall argument to be free of threats to reliable learning.

The Only Correct Interpretation of the Data: It is in the Folding

An important mode of learning from error is to consider why certain experiments are incapable of telling experimenters what they want to know; and this case offers several illustrations. By mixing synthetic versions of the two proteins together in a test tube, they were able to convert common prions (PrP-C) into scrapie prions (PrP-Sc) (in vitro), but they did not understand what was actually causing
the switch. Moreover, they knew they did not, and they knew something about why. The infectious form has the same amino acid sequence as the normal type: studying the amino acid sequence does not enable us to reveal what made the difference. If exceptions to the ‘central dogma’ were precluded, and it is assumed that only nucleic acid directs replication of pathogens, there was no other place to look. But what if transmission by pathogenic proteins occurs in a different way?

Maybe the difference is in the way the protein is folded. Researchers hypothesized that the scrapie protein propagates itself by getting normal prions to flip from their usual shape to a scrapie shape. This would explain ‘transmission’ without nucleic acid, and sure enough they were able to replicate such flipping in vitro. But understanding the mechanism of pathological folding required knowing something about the structures of common as opposed to scrapie prions. A big experimental obstacle was not being able to discern the prion’s three-dimensional structure at the atomic level. Exploiting the obstacle provided the key.

### Magic Angle Spinning: Exploiting an Obstacle

The central difference between normal and pathogenic prions permits the normal but not the abnormal prion to have its structure discerned by known techniques, e.g. nuclear magnetic resonance (NMR) for solutions: The normal form, PrP-C, is soluble; PrP-Sc is not.

NMR spectroscopy provides an image of molecular structure: Put a material inside a very high magnetic field, hit it with targeted radio waves, and its particles react to reveal their structure. But it will not work for clumpy scrapie prions, PrP-Sc. Maybe solid-state NMR could detect them?

Even so, they would need trillions of molecules to get a signal – amplification – but this would also amplify the interference of neighbouring molecules in the non-soluble PrP-Sc. They want to find out what it would be like if they were able to make it soluble, even though they cannot literally do so. They need to amplify to get a signal, but also somehow subtract the interference of neighbouring molecules. Here is where ‘magic angle spinning’ enters.

*The Magic is to erase the influence of these neighbouring molecules.*

If the sample, crushed into a powder form, spins within a magnetic field at a special angle to that field – 54.7 degrees – the influence of a molecule’s environment is cancelled out. The effect on the spectrum from the magnetic interactions between the molecules vanishes. Knowledge of the magic angle stems from quantum mechanics, but that it works to negate interactions is shown with known molecular structures.

While the molecules cannot all be lined up at the magic angle, they can, *on average*, if they are spun fast enough (with respect to every other molecule in the sample). In this we have one of the deepest and least appreciated aspects of the role of models of relative frequencies, and intermediate statistical mod-
the experimental inference that is licensed fails to provide evidence for the primary hypothesis

the experimental hypothesis has not passed a severe test

the assumptions of the experiment are not met sufficiently by the actual data.

With transgenic mice designed to produce tons of normal prions, with synthetic prions, with protein folding in vitro, and so on, experimenters create a series of experiments $E_1, E_2, E_3, E_4, \ldots, E_n$ that enable arguments from error to stand. The immediate goal in each is not to rule out rival theories, but to rule out a mistaken interpretation of results from this experiment.

But what do the pieces say about actual (real-world) prion disease? Might these be relevant only for Frankenstein mice, yeast, synthetic prions and mimicked scrapie in a test tube? To combine these pieces requires understanding the errors or distortions that remain, and those being avoided or subtracted out. (What's being amplified? What's being silenced?) We can investigate and learn about these limits and distortions by deliberately amplifying and controlling them in known or canonical cases.

In so doing we capitalize on experimental benchmarks, calibration standards, a variety of extrapolation models and a suite of relevant repertoires of errors. By the time magic angle spinning is used to learn about the structure of prions, it is a well-understood and reliable instrument. Imagining that an instrument works only when it is used on a known sample is to imagine that it can read our minds, and that it conspires to trick us just when we are faced with unknown samples. This would be like my claiming that all the scales have conspired to show that I have gained weight when I have not, while allowing that the scales work fine with objects of known weight. If any of the scales were faulty, this would show up when they were used on objects of known weight. We need only deny that a mysterious power conspires to make all observations fit our deductions on known samples, but not on unknown ones. That would be a radical obstacle to learning.

The powdered PrP-Sc spun around in magic angle spinning turbines can tell us about the three-dimensional structure of PrP-Sc because we know how the turbines work with known solid-state specimens. While only the structural relationship is discerned, that is all we need for this piece of the puzzle. Other interlocking hypotheses make the pieces relevant to what we want to know. To understand specifically how and why this works we should consider how each experiment checks, unearths, amplifies and erases the threats or shortcomings of others.

Even today, the prion evidence can be accommodated by those who still maintain there could be some nucleic acid locked within the prion protein; indeed, they revise and develop their alternatives to deliberately account for the known evidence. It was only after decades of accumulated growth of experimental knowledge that defenders of the older paradigm could construct alternatives such as virions (a virus wrapped in a protein) that could explain why no protein is discerned, despite years of trying. Nonetheless, the idea that transmission is through protein misfolding is granted even by the alternative virion theories, as I understand them. As such, work on developing rivals does not really hamper progress (though it frustrates 'prion-only' theorists); and the corroboration of the general prion transmission principle is not compromised. In general, the development of rivals to a hypothesis $H$ that could account for existing data plays an important role in experimental learning, even where it must be denied that the data are evidence for the rivals constructed. It is a way for defenders of $H$ to uncover their own errors, and for sceptics about $H$ to show which aspects have so far not been well-tested. With well-grounded experimental knowledge, these rivals do not alter the well-testedness of the aspects severely probed: they have their own lives.

Scientists are rarely fully explicit about or even aware of why their methods and strategies work when they do. Providing such an illumination would be an important task for a future philosophy of experiment: the key is to unearth how experiments manage to free themselves from threats of error.
3 Mayo, ‘Learning from Error: How Experiment Gets a Life (of its Own)


2. If there is a need to restore the more usual distinction between experimental and observational research, the former might be dubbed ‘manipulative experiment’ and the latter ‘observational experiment’.

3. In Mayo, Error and the Growth of Experimental Knowledge, errors number 4 and 5 were collapsed under the rubric of ‘experimental assumptions’.


6. The transmissibility of TSEs was accidentally demonstrated in 1937, when a population of Scottish sheep was inoculated against a common virus with a formalin extract of brain tissue unknowingly derived from an animal with scrapie. After two years, nearly 10 per cent of the flock developed scrapie. Scrapie was subsequently transmitted experimentally to sheep.


10. Actually, the very constraints that Kuhn accords to ‘normal science’ or testing within a paradigm are at the heart of what gives stability to experimental effects. As I argue in Chapter 2 of Mayo, Error and the Growth of Experimental Knowledge (some may consider this a radical ‘deconstruction’ of Kuhn), normal science can be seen to describe the highly constrained, local experimental probing. Taking the stringent demands of normal science seriously forces the experimenter to face the music, to admit he has not solved his problem rather than seek to conveniently change the problem. Where Kuhn errs is in throwing away the stringent demands he recognizes for probing within a paradigm when it comes to theory change. Moreover, Kuhn gives no argument for postulating a radical discontinuity between the role of evidence in conducting normal tests, and in finding flaws in a large paradigm (‘revolutionary’ science).

11. There is a corresponding argument that H is an incorrect construal of data. These arguments from error can equivalently be put in terms of inferring H (or not-H) with severity.


16. Popper once wrote to me that he regretted not having learned modern statistical methods. If he had, he might have seen error probabilities as offering the third way he sought. That was clearly Peirce’s way, although he anticipated modern statistics.


23. Viewing evidential appraisal as unable to live a life apart from cost benefit analysis is a conception I reject. It has done much damage in evidence-based policy: if disagreements about interpreting data are indistinguishable from disagreements about matters of subjective opinions and values, then each side gets its own scientific experts, often resulting in ‘junk science’ all around. We can ask should data on the frequency of BSE/rmad cow disease be interpreted according to the economic values of the beef industry? Probably not, which is why obtaining sound inferences needs to be distinct from, and not inextricably intertwined with, subsequent policy decisions.

24. Hacking, Representing and Intervening.

25. One can depict the test in myriad ways. The artefact error might be seen as standing in for a null hypothesis; it asserts the presence of a given error.
26. But it was also the basis for constructing possible alternatives, such as virions.


28. Magic angle spinning was invented by physicist E. Raymond Andrew. For a discussion, see [http://www.magnet.fsu.edu/education/tutorials/tools/probes/magicangle.html](http://www.magnet.fsu.edu/education/tutorials/tools/probes/magicangle.html) [accessed 19 October 2013].

29. Nowadays, researchers know how to perform what is called cyclical amplification: lopping off the ends of the pathogenic (beta) helix, they get huge amounts of PrP-Sc starting with minute quantities. This affirms a correct understanding of the misfolding mechanism (pathogenic prions propagate through misfolding at the ends of the helix) and also gives an important new tool to detect prion disease in living animals for the first time.

30. Hacking, *Representing and Intervening*.
