An Ad Hoc Save of a Theory of Adhocness? Exchanges with John Worrall

Deborah G. Mayo

In large part, the development of my concept of severity arose to deal with long-standing debates in philosophy of science about whether to require or prefer (and even how to define) novel evidence (Musgrave, 1974, 1989; Worrall 1989). Worrall's contribution represents the latest twists on our long-running exchange on the issue of novel evidence, beginning approximately twenty years ago; discussions with Musgrave around that time were also pivotal to my account.[1] I consider the following questions:

1. *Experimental Reasoning and Reliability*: Do distinct uses of data in science require distinct accounts of evidence, inference, or testing?
2. *Objectivity and Rationality*: Is it unscientific (ad hoc, degenerating) to use data in both constructing and testing hypotheses? Is double counting problematic only because and only when it leads to unreliable methods?
3. *Metaphilosophy*: How should we treat counterexamples in philosophical arguments?

I have argued that the actual rationale underlying preferring or requiring novel evidence is the intuition that it is too easy to arrive at an accordance between nonnovel data and a hypothesis (or model) even if H is false: in short the underlying rationale for requiring novelty is severity. Various impediments to severity do correlate with the double use of data, but this correlation is imperfect. As I put it in Mayo (1996): "Novelty and severity do not always go hand in hand: there are novel tests that are not severe and severe tests that are not novel. As such, criteria for good tests that are couched in terms of novelty wind up being either too weak or too strong, countenancing poor tests and condemning excellent ones" (p. 253). This

[1] Mayo (1991; 1996, p. xv).

malady is suffered by the version of novelty championed by Worrall, or so I argue. His chapter turns us directly to the problem of metamethodology with respect to a principle much debated in both philosophy of science and statistical practice.

The UN requirement – or, as he playfully calls it, the UN Charter – is this:

1.1 Use-Novelty Requirement (UN Charter): For data **x** to support hypothesis H (or for **x** to be a good test of H), H should not only agree with or "fit" the evidence **x**, **x** must itself *not have been used* in H's construction.

For example, if we find data **x** anomalous for a theory or model, and we use **x** to arrive at a hypothesized explanation for the anomaly $H(\mathbf{x})$, it would violate the UN Charter to also regard **x** as evidence for $H(\mathbf{x})$. Much as with the rationale for varying the evidence (see also Chapter 2, Exchanges with Chalmers), use-novelty matters just to the extent that its violation inhibits or alters the reliability or stringency of the test in question. We can write the severity requirement in parallel to the UN Charter:

1.2 Severity Requirement: For data **x** to support hypothesis H (or for **x** to be a good test of H), H should not only agree with or "fit" the evidence **x**, H must have passed a stringent or severe test with **x**.

If UN violations alter a test's probativeness for the inference in question, the severity assessment must be adjusted accordingly.

However, as I have argued, some cases of "use-constructed" hypotheses succeed in being well tested by the same data used in their construction. To allude to a case discussed in Mayo (1996, p. 284), an apparent anomaly for the General Theory of Relativity (GTR) from the 1919 Sobral eclipse results was shown to be caused by a mirror distortion from the sun's heat. Although the eclipse data were used both to arrive at and to test $A(\mathbf{x})$ – the hypothesized mirror distortion explanation – $A(\mathbf{x})$ passed severely because it was constructed by a reliable or stringent rule. Because Worrall agrees with the severity goal, these cases stand as counterexamples to the UN requirement. Worrall's chapter discusses his recent attempts to accommodate these anomalies; surprisingly, Worrall is prepared to substantially adjust his account of confirmation to do so.

In particular, he allows my counterexamples to stand, but regards them as involving a distinct kind of support or corroboration, a kind he has developed to accomodate UN violations. For instance, if $A(\mathbf{x})$ is use-constructed to account for **x** which is anomalous for theory T, then the inference to $A(\mathbf{x})$ gets what Worrall calls "conditional support." By this he means that $A(\mathbf{x})$ is legitimately inferred only conditional on already assuming T, the

theory to be "saved." So for Worrall the UN requirement still stands as necessary (and sufficient) for full-bodied support, but data **x** may still count as evidence for use-constructed $A(\mathbf{x})$ so long as we add "conditional on accepting an overarching theory *T*."

But I do not see how Worrall's attempt to save the UN Charter can adequately accommodate the counterexamples I have raised. It is false to suppose it is necessarily (or even commonly) the case that use-constructed hypotheses assume the truth of some large-scale theory, whether in the case of blocking an anomaly for theory *T* or in any of the other use-constructions I have delineated (Mayo, 1996, 2008). Certainly using the eclipse data to pinpoint the source of the GTR anomaly did not involve assuming the truth of GTR. In general, a use-constructed save of theory *T* takes the form of a hypothesis designed to block the alleged anomaly for *T*.

1.3 A use-constructed block of an anomaly for *T*:

$A(\mathbf{x})$: the anomalous data **x** are due to factor *F*, not the falsity of *T*, or
$A(\mathbf{x})$ explains why the data **x** did not accord with the predictions from *T*.

By lumping together all cases that follow this logical pattern, Worrall's account lacks the machinery to distinguish reliable from unreliable use-constructions. As a result, I argue, Worrall's account comes up short when it comes to real experimental inferences:

> Any philosophy of experimental testing adequate to real experiments must come to grips with the fact that the relationship between theory and experiment is not direct but is mediated along the lines of the hierarchy of models and theories.... At various stages of filling in the links, it is standard to utilize the same data to arrive at as well as warrant hypotheses.... As a matter of course, then, the inferences involved violate even the best construals of the novelty requirement. (Mayo, 1996, p. 253)

By maintaining that all such use-constructions are conditional on already assuming the truth of some overarching theory or "research program," Worrall's philosophy is redolent of the image of scientists as locked into "theory-laden" paradigms (Lakatos, Kuhn). Conversely, by regarding UN as sufficient for warranted inference, Worrall overlooks the fact "that there is as much opportunity for unreliability to arise in reporting or interpreting (novel) results given knowledge of theoretical predictions as there is... in arriving at hypotheses given knowledge of (non-novel) results" (ibid, p. 254). By recognizing that what matters is the overall severity with which a claim may be inferred, we have a desideratum that allows us to

discriminate, on a case-by-case basis, whether UN violations matter and, if so, how we might correct for them.

I begin by discussing Worrall's treatment of use-construction in blocking anomalies and then turn to some confusions and errors that lead Worrall to forfeit a discriminatory tool that would seem to fulfill the (Popperian) testing philosophy that he himself endorses.

2 Use-Constructing in Blocking Anomalies: Must All of *T* Be Assumed?

We are all familiar with a variety of "rigging" procedures so as to accommodate data while protecting pet hypotheses rather than subjecting them to scrutiny. One of Worrall's favorite examples is Velikovsky's method for use-constructing hypotheses to save his theory when confronted with any anomalous data **x**:

> The lack of records in cultures $C_1, \ldots, C_n$ and their (arguable) presence in $C'_1, \ldots, C'_m$ gives very good reason for holding the specific collective-amnesia version of Velikovsky's theory that he proposed *if* you already hold Velikovsky's general theory, *but* (and this is where the initial UN intuitions were aimed) those data give you absolutely no reason at all for holding that general theory in the first place. (Worrall, this chapter, pp. 136–7)

But why suppose that the inference to blocking a T anomaly assumes all of T? We know that, even if it is warranted to deny there is evidence against T, this fact alone would not provide evidence *for* T, and there is no reason to saddle every use-constructed save with committing so flagrant a fallacy (circularity). Obviously any method that assumes T in order to save T is minimally severe, but it is false to suppose that in use-constructing $A(\mathbf{x})$, T is assumed. It is not even clear why accepting Velikovsky means that any lack of records counts as evidence for the amnesia hypothesis, unless it is given that no other explanation can exist for the anomaly, as I take it Worrall does (note 10, p. 136). But are we to always imagine this? I put this aside. Even a proponent of Velikovsky's dodge could thwart Worrall's charge as follows.

> *V-dodger*: I am not claiming that lack of records of the cataclysms described in my theory T is itself evidence for T (other records and considerations provide that); I am simply saying that I have a perfectly sound excuse, $A(\mathbf{x})$, for discounting the apparent anomaly for my theory.

Despite the ability to escape Worrall's charge, the flaw in the V-dodger's inference seems intuitively obvious. The severity account simply provides

some systematic tools for the less obvious cases. We are directed to consider the use-construction rule R leading from $\mathbf{x}$ to the inference $A(\mathbf{x})$ and the associated threats of error that could render the inference unwarranted. Here, we can characterize the rule R in something like the following manner.

Rule R (Velikovsky's scotoma dodge): For each possible set of data $\mathbf{x}^i$ indicating that culture C^i has no records of the appropriate cataclysmic events, infer $A^i(\mathbf{x}^i)$: culture C^i had amnesia with regard to these events.

The blocking hypothesis $A^i(\mathbf{x}^i)$ is use-constructed to fit data $\mathbf{x}^i$ to save Velikovsky from anomaly.

Clearly, rule R prevents any observed anomaly of this form to threaten Velikovsky's theory, even if the culture in question had not suffered amnesia in the least. If one wanted to put this probabilistically, the probability of outputting a Velikovsky dodge in the face of anomaly is maximal, even if the *amnesia explanation* is false (a case of "gellerization") – therefore, severity is minimal. Because rule R scarcely guards against the threat of erroneously explaining away anomalies, we would say *of any particular output of rule R* that the observed fit fails to provide evidence for the truth of $A(\mathbf{x}_0^i)$.

Notice that one need not rule out legitimately finding evidence that a given culture had failed to record events that actually occurred, whether due to memory lapses, sloppy records, or perhaps enforced by political will. For example, we could discern that all the textbooks in a given era were rewritten to expunge a given event, whose occurrence we can independently check. But with Velikovky's rule R there is no chance that an erroneous attribution of scotoma (collective amnesia) would be detected; nothing has been done that could have revealed this fact, at least by dint of applying rule R.

Although we condemn inferences from tests that suffer from a low probability of uncovering errors, it is useful to have what I call "canonical errors" that stand as extreme cases for comparison (cases of zero severity). Velikovsky's case gives one. We utterly discredit any inference to $A(\mathbf{x})$ resulting from Velikovsky's use-construction rule, as seems proper. It is surprising, then, that Worrall's account appears to construe Velikovsky's gambit as no worse off than any other use-constructed saves, including those that we would consider altogether warranted.

The detailed data analysis of eclipse plates in 1919 warranted the inference that "the results of these (Sobral Astrographic) plates are due to systematic distortion by the sun and not to the deflection of light" (Mayo, 1996, p. 284). To warrant this explanation is to successfully block an interpretation

of those data as anomalous for GTR. In Worrall's account, however, all use-constructed saves of theory T are conditional on assuming T; the only way they can avoid being treated identically to the case of Velikovsky's dodge is if *other*, independent support arises for accepting T.

As a matter of fact, however, the data-analytic methods, well-known even in 1919, did not assume the underlying theory, GTR, nor is it correct to imagine Eddington arguing that, provided you accept GTR, then the mirror distortion due to the Sun's heat explains why the 1919 Sobral eclipse results were in conflict with GTR's predicted deflection (and in agreement with the Newtonian prediction). GTR does not speak about mirror distortions. Nor were even the staunchest Newtonians unable to agree (not that it was immediately obvious) that the detailed data analysis showed that unequal expansion of the mirror caused the distortion. It was clear the plates, on which the purported GTR anomaly rested, were ruined; accepting GTR had nothing to do with it. Nor could one point to GTR's enjoying more independent support than Newton at the time – quite the opposite. (Two data sets from the same eclipse afforded highly imprecise accordance with GTR, whereas Newton enjoyed vast support.) Nor would it make sense to suppose that vouchsafing the mirror distortion depended on waiting decades until GTR was warranted, as Worrall would seem to require.

Thus, I remain perplexed by Worrall's claim that we need "to recognise just how conditional (and *ineliminably* conditional) the support at issue is in all these cases" (this volume, p. 135). By this, he means not that there are assumptions – because that is always true, and Worrall is quite clear he does not wish to label all cases as giving merely conditional support. He means, rather, that the entire underlying theory is assumed. We have seen this to be false.

My goal (e.g., in Mayo, 1996, sec. 8.6) was to illustrate these counterexamples to the UN requirement, at several stages of testing:

> The arguments and counterarguments [from 1919 to ~1921] on both sides involved violating UN. What made the debate possible, and finally resolvable, was that all... were held to shared criteria for acceptable and unacceptable use-constructions. It was acceptable to use any evidence to construct and test a hypothesis... so long as it could be shown that the argument procedure was reliable or severe. (p. 289)

Although the inferences, on both sides of the debate, strictly violated UN, they were deliberately constrained to reflect what is correct, at least approximately, regarding the cause of the anomalous data.

These kinds of cases are what led me to abandon the UN Charter, and Worrall has yet to address them. Here the "same" data are used both to identify and to test the source of such things as a mirror distortion, a plane crash, skewed data, a DNA match, and so on – without threats from uncertain background theories ("clean tests"). In statistical contexts, the stringency of such rules may be quantitatively argued:

A Stringent Use-Construction Rule (*R*-α): The probability is very small, $1 - \alpha$, that rule *R* would output $H(\mathbf{x})$ unless $H(\mathbf{x})$ were true or approximately true of the procedure generating data **x**. (Mayo, 1996, p. 276)

Once the construction rule is applied and a particular $H(\mathbf{x}_0)$ is in front of us, we evaluate the severity with which $H(\mathbf{x}_0)$ has passed by considering the stringency of the rule *R* by which it was constructed, taking into account the particular data achieved. What matters is not whether *H* was deliberately constructed to accommodate **x**; what matters is how well the data, together with background information, rule out ways in which an inference to *H* can be in error.

2.1 Deducing a Version (or Instantiation) of a Theory

At several junctures, it appears that Worrall is taking as the exemplar of a UN violation "using observational data as a premise in the deduction of some particular version of a theory" (p. 131) so that there is virtually no threat of error. True, whenever one is in the context wherein all of the givens of Worrall's inference to "the representative or variant of the theory" are met, we have before us a maximally severe use-construction rule (α would equal 1). We can agree with his claim that "[w]e do want to say that **x** supports $T(\mathbf{x})$ in some quite strong sense," (see p. 151), where $T(\mathbf{x})$ is what he regards as the variant of theory *T* that would be instantiated from the data **x**. Confronted with a particular $T(\mathbf{x}_0)$, it would receive maximal support – provided this is understood as inferring that $T(\mathbf{x}_0)$ is the variant of *T* that would result if *T* were accepted and $\mathbf{x}_0$ observed. Instantiating for the wave theory, *W*, Worrall asserts: "$\mathbf{x}_0$ definitely supports $W(\mathbf{x}_0)$ in the conditional sense in that it establishes $W(\mathbf{x}_0)$ as *the* representative of the general theory *W* if that theory is to work at all" (replacing *e* with $\mathbf{x}_0$).[2] (p. 152) Although this

[2] An example might be to take the results from one of the GTR experiments, fix the parameter of the Brans-Dicke theory, and infer something like: if one were to hold the B-D theory, then the adjustable constant would have to be such-and-such value, for example, $q = 500$. (See Chapter 1, Section 5.2.)

inference is not especially interesting, and I certainly did not have this in mind in waging the counterexamples for the UN Charter, handling them presents no difficulty. If the assumptions of the data are met, the "inference" to the instantiation or application of theory T is nearly tautological.

The question is why Worrall would take this activity as his exemplar for use-constructed inferences in science. Certainly I would never have bothered about it if that was the sort of example on which the debate turned. Nor would there be a long-running debate in methodological practice over when to disallow or make adjustments because of UN violations and why. Yet, by logical fiat – construing all UN violations as virtually error-free inferences that aspire to do no more than report a specific variant of a theory that would fit observed data – the debate is settled, if entirely trivialized. If philosophers of science are to have anything useful to say about such actual methodological debates, the first rule of order might be to avoid interpreting them so that they may be settled by a logical wand.

Worrall claims to have given us good reasons for accepting his account of confirmation in the face of anomalies – where the anomalies are counterexamples to his view that UN is necessary for full-bodied confirmation. We might concur, in a bit of teasing reflexivity, that Worrall has given reason to support his handling of anomalies if you already hold his account of conditional support! But I doubt he would welcome such self-affirmation as redounding to his credit. This point takes me to a cluster of issues I place under "metaphilosophy."

3 Metaphilosophy: The Philosophical Role of Counterexamples

To a large extent, "the dispute between those who do and those who do not accept some version of the novelty principle emerges as a dispute about whether severity – or, more generally, error characteristics of a testing process – matters" (Mayo, 1996, p. 254). If it is assumed that whether H is warranted by evidence is just a function of statements of evidence and hypotheses, then it is irrelevant how hypotheses are constructed or selected for testing (I call these evidential-relation accounts). What then about the disagreement even among philosophers who endorse something like the severity requirement (as in the case of Worrall)? Here the source of disagreement is less obvious, and is often hidden: to dig it up and bring it to the surface requires appealing to the philosopher's toolkit of counterexamples and logical analysis. However, "philosopher's examples" are anything but typical, so one needs to be careful not to take them out of their intended context – as counterexamples!

3.1 Counterexamples Should Not Be Considered Typical Examples: The SAT Test

Now Worrall agrees with the general severity rationale: "the underlying justification is exactly the same as that cited by Mayo in favour of her own approach . . . a theory T is supported in this [strong] sense by some evidence e only if (and to the extent that) e is the outcome (positive so far as T is concerned) of some severe test of T" (Worrall, this volume, p. 144). We concur that, for a passing result to count as severe, it must, first of all, *be a passing result*; that is, the data must fit or accord with hypothesis H (where H can be any claim). Although Worrall often states this fit requirement as entailment, he allows that statistical fits are also to be covered. The key difference regards *what more* is required to warrant the inference to H. Should it be Worrall's UN criterion, or my severity criterion?

To argue for the latter, my task is to show how UN could be violated while intuitively severity is satisfied. In this I turned to the usual weapon of the philosopher: counterexamples. Observe what happens in cases where it is intuitively, and blatantly, obvious that a use-constructed hypothesis is warranted: the method that uses the data to output $H(\mathbf{x})$ is constrained in such a way that $H(\mathbf{x})$ is a product of what is truly the case in bringing about data $\mathbf{x}$. Worrall and like-minded use-novelists often talk as if an accordance between data and hypothesis can be explained in three ways: it is due to (1) chance, (2) the "blueprint of the universe" (i.e., truth or approximate truth of H), or (3) the ingenuity of the constructor (Worrall, 1989, p. 155). That hypotheses can be use-constructed reliably is precisely what is overlooked. In my attempts to lead them to the "aha" moment, I – following the philosopher's craft – sought extreme cases that show how use-constructed hypotheses can pass with high or even maximal severity; hence, the highly artificial example of using the data on the SAT scores to arrive at the mean SAT score. As I made clear, "the extreme represented by my SAT example was just intended to set the mood for generating counterexamples" (Mayo, 1996, p. 272), after which I turn to several realistic examples. Worrall (who is not alone) focuses on the former and gives little or no attention to the latter, realistic cases.

Ironically, it was Musgrave's reaction long ago to such flagrant cases that convinced me the Popperians had erred in this manner: "An older debt recalled in developing the key concept of severe tests is to Alan Musgrave" (Mayo, 1996, p. xv). Actually, as Musgrave reminds me, the example that convinced him was the incident that first convinced me that UN is not necessary for a good test: using data on the dent in my Camaro to hunt for

a car with a tail fin that would match the dent, to infer that "it is practically impossible for the dent to have the features it has unless it was created by a specific type of car tail fin" (p. 276). The point is that counterexamples serve as this kind of tool in philosophy, and no one would think the user of the counterexamples intended them as typical examples. Yet some charge that I must be regarding the SAT averaging as representative of scientific hypotheses, forgetting that it arises only in the service of getting past an apparent blind spot.

We should clear up a problem Worrall has with the probabilistic statement we make. He considers the example of deducing $H(\mathbf{x})$ from data $\mathbf{x}$ (e.g., deducing the average SAT score from data on their scores). The probability $H(\mathbf{x})$ would be constructed, if in fact the data came from a population where $H(\mathbf{x})$ is false, is zero. (Because this is true for any $\mathbf{x}$, it is also true for any instance $\mathbf{x}_0$). But Worrall claims it would be undefined because the denominator of a conditional probability of a false claim is zero. Now the correct way to view an error-probabilistic statement, for example,

$$\text{P(test } T \text{ outputs } H(\mathbf{x});\ H(\mathbf{x}) \text{ is false)},$$

is *not* as a conditional probability but rather a probability *calculated under the assumption that* $\mathbf{x}$ *came from a population where* $H(\mathbf{x})$ *is false.* The probability that a maximally severe use-construction rule outputs $H(\mathbf{x}_0)$, calculated under the assumption that $H(\mathbf{x}_0)$ is false, is zero – not undefined. Moreover, if we bar conditional probabilities on false hypotheses, then Bayesians could never get their favorite theorem going because they must exhaust the space of hypotheses.

3.2 Equivocations and Logical Flaws

If counterexamples will not suffice (in this case, to deny UN is necessary for severity), a second philosophical gambit is to identify flaws and equivocations responsible for leading astray even those who profess to share the goal (severity). *But one can never be sure one has exhausted the sources of confusions!* Worse is that the analytic labors carefully crafted to reveal the logical slip can give birth to yet new, unintended confusions. This seems to have happened here, and I hope to scotch it once and for all.

Everything starts out fine: Worrall correctly notes that I identify, as a possible explanation for the common supposition that UN is necessary for severity, a slippery slide from a true assertion – call it (a) – to a very different assertion (b), which need not be true:

(a) A use-constructed procedure is guaranteed to output an $H(\mathbf{x})$ that fits $\mathbf{x}$, "no matter what the data are."

(b) A use-constructed procedure is guaranteed to output an $H(\mathbf{x})$ that fits $\mathbf{x}$, "no matter whether the use-constructed $H(\mathbf{x})$ is true or false" (Mayo, 1996, p. 270; Worrall, this volume, p. 148).

Giere, for example, describes a scientist unwilling to consider any model that did not yield a prediction in sync with an observed effect $\mathbf{x}$. "Thus we know that the probability of any model he put forward yielding [the correct effect $\mathbf{x}$] was near unity, independently of the general correctness of that model" (Giere, 1983, p. 282). It is this type of multiply ambiguous statement, I argue, that leads many philosophers to erroneously suppose that use-constructed hypotheses violate severity. Pointing up the slide from (true) assertion (a) to (false) assertion (b) was intended to reveal the equivocation. Let me explain.

A use-constructed test procedure has the following skeletal form:

Use-Constructed Test Procedure: Construct $H(\mathbf{x})$ to fit data $\mathbf{x}$; infer that the accordance between $H(\mathbf{x})$ and $\mathbf{x}$ is evidence for inferring $H(\mathbf{x})$.

We write this with the variable $\mathbf{x}$, because we are stating its general characterization. So, *by definition*, insofar as a use-constructed procedure is successfully applied, it uses $\mathbf{x}_0$ to construct and infer $H(\mathbf{x}_0)$, where $\mathbf{x}_0$ fits $H(\mathbf{x}_0)$. This is captured in assertion (a). But assertion (a) alone need not yield the minimally severe test described in assertion (b); it need not even lead to one with low severity. The construction rule may ensure that false outputs are rare. We may know, for example, that anyone prosecuted for killing JonBenet Ramsey will have to have matched the DNA from the murder scene; but this is a reliable procedure for outputting claims of form:

The DNA belongs to Mr. X.

In any specific application it outputs $H(\mathbf{x}_0)$, which may be true or false about the source of the data, but the probability that it outputs false claims is low. The familiar argument that use-constructed tests are invariably minimally severe, I suggest, plays on a (fallacious) slide from assertion (a) to assertion (b).

Having gotten so used to hearing the Popperian call for falsification, it is sometimes forgotten that his call was, strictly speaking, for falsifying hypotheses, *if false.* Admittedly, Popper never adequately cashed out his severity idea, but I would surmise that, if he were here today, he would agree that some construction procedures, although guaranteed to output some

H(**x**) *or other*, whatever the data, nevertheless ensure false outputs are rare or even impossible.

Worrall sets out my argument with admirable clarity. Then something goes wrong that numerous exchanges have been unable to resolve. His trouble is mainly as regards claim (a). Now claim (a) was intended to merely capture what is generally assumed (*by definition*) for any use-constructed procedure. So, by instantiation, claim (a) holds for the examples I give where a use-constructed procedure yields a nonsevere test. From this, Worrall supposes that claim (a) is necessary for nonseverity, but this makes no sense. Were claim (a) required for inseverity, then violations of claim (a) would automatically yield severity. Then hypotheses that do not even fit the data would automatically count as severe! But I put this error aside. More egregiously, for current purposes, he argues that claim (a) is false! Here is where Worrall's logic goes on holiday.

He considers Velikovsky's rule for blocking anomalies by inferring that the culture in question suffered amnesia *A*(**x**). Worrall says, consider a specific example of a culture – to have a concrete name, suppose it is the Thoh culture – and suppose no records are found of Velikovsky-type events. Velikovsky conveniently infers that the apparent anomaly for his theory is explained by amnesia:

A(Thoh): Thoh culture suffered from amnesia (hence no records).

Says Worrall, "It seems, then, to be straightforwardly untrue that a successful fit between" *A*(Thoh) and **x** "is assured no matter what [**x**] is" (p. 149). Quite so! (For instance, the procedure would not output *A*(Thoh), or any claims about the Thoh culture, if the observation was on some other culture). But this does not show that assertion (a) is false. It could only show that assertion (a) is false by an erroneous instantiation of the universal claim in assertion (a). The assertion in (a) is true because every anomalous outcome will fit *some Velikovsky dodge or other.* It does not assert that all anomalous cultures fit a *particular* instantiation of the Velikovsky dodge, e.g., *A*(Thoh).[3]

[3] Worrall seems to reason as follows:

1. According to assertion (a), for any data **x**, if **x** is used to construct *A*(**x**), then **x** fits *A*(**x**).
2. But suppose the data from the Thoh culture is used to construct *A*(Thoh).
3. (From assertion (a) it follows that) all data **x** would fit *A*(Thoh) (i.e., a successful fit between *A*(Thoh) and **x** "is assured no matter what **x** is").

Then from the falsity of premise 3, Worrall reasons that premise (a) is false. But premise 3 is an invalid instantiation of the universal generalization in premise (a)! It is unclear whether Worrall also takes this supposed denial of assertion (a) as denying assertion (b), but to do so is to slip into the fallacy that my efforts were designed to avoid. (For further discussion of variations on this fallacy, e.g., in Hitchcock and Sober, 2004, see Mayo, 2008).

Sometimes a gambit that a philosopher is sure will reveal a logical flaw instead creates others. Pointing up the faulty slide from the truth of assertion (a) to that of assertion (b) was to have illuminated the (false but common) intuition that UN is necessary for severity. Instead we have been mired in Worrall's resistance to taking assertion (a) as true for use-constructed procedures – something I took to be a matter of mere definition, which just goes to show that one cannot always guess where the source of difficulties resides. Hopefully now no obstacles should remain to our agreement on this issue.

4 Concluding Comment on the Idea of a Single Account of Evidence (Remarks on Chapters 2, 3, and 4)

I do not claim that all of science involves collecting and drawing inferences from evidence, only that my account is focused on inference. As varied as are the claims that we may wish to infer, I do not see that we need more than one conception of what is required for evidence to warrant or corroborate a claim. Worrall berates me for holding a "one-size-fits-all" account of inference that always worries about how well a method has probed for the errors that threaten the inference. Similar sentiments are voiced by Chalmers and Musgrave. Granted data may be used in various ways, and we want hypotheses to be not just well tested, but also informative; however, if we are talking about the warrant to accord a given inference, then I stand guilty as charged.

I cannot really understand how anyone could be happy with their account of inference if it did not provide a unified requirement. In the context of this chapter, Worrall's introduction of "conditional evidence" was of no help in discriminating warranted from unwarranted use-constructions. The severity desideratum seems to be what matters. Similarly, Chalmers's "arguments from coincidence" and Musgrave's "inference to the best tested portion of an explanation" in Chapters 2 and 3, respectively, are all subsumed by the severity account. Different considerations arise in *applying* the severity definition, and different degrees of severity are demanded in different cases, but in all cases the underlying goal is the same. The whole point of the approach I take is to emphasize that what needs to have been probed are the threats of error in the case at hand. Even if one adds decision-theoretic criteria, which we will see leads Laudan to argue for different standards of evidence (Chapter 9), my point is that, *given the standards*, whether they are satisfied (by the data in question) does not change.

The deepest source of the disagreements raised by my critics, I see now, may be located in our attitudes toward solving classic problems of

evidence, inference, and testing. The experimental account I favor was developed precisely in opposition to the philosophy of science that imagines all inferences to be paradigm-laden in the sense Kuhnians often espouse, wherein it is imagined scientists within paradigm T_1 circularly defend T_1 against anomaly and have trouble breaking out of their prisons. In this I am apparently on the side of Popper, whereas Musgrave, Chalmers, Worrall (and Laudan!) concede more to Lakatos and Kuhn. It is to be hoped that current-day Popperians move to a position that combines the best insights of Popper with the panoply of experimental tools and methods we now have available.

At the same time, let me emphasize, there are numerous gaps that need filling to build on the experimentalist approach associated with the error-statistical account. The example of use-novelty and double counting is an excellent case in point. Although in some cases, understanding the way formal error probabilities may be altered by double counting provides striking illumination for entirely informal examples, in other cases (unfortunately), it turns out that whether error probabilities are or should be altered, even in statistics, is unclear and requires philosophical–methodological insights into the goals of inference. This is typical of the "two-way street" we see throughout this volume. To help solve problems in practice, philosophers of science need to take seriously how they arise and are dealt with, and not be tempted to define them away. Conversely, in building the general experimentalist approach that I label the error-statistical philosophy of science, we may at least find a roomier framework for re-asking many philosophical problems about inductive inference, evidence and testing.

References

Giere, R.N. (1983), "Testing Theoretical Hypotheses," pp. 269–98 in J. Earman (ed.), *Testing Scientific Theories*, Minnesota Studies in the Philosophy of Science, vol. 10, University of Minnesota Press, Minneapolis.

Hitchcock, C., and Sober, E. (2004), "Prediction Versus Accommodation and the Risk of Overfitting," *British Journal for the Philosophy of Science*, 55: 1–34.

Mayo, D.G. (1991), "Novel Evidence and Severe Tests," *Philosophy of Science*, 58: 523–52.

Mayo, D.G. (1996), *Error and the Growth of Experimental Knowledge* (Chapters 8, 9, 10), University of Chicago Press, Chicago.

Mayo, D.G. (2008), "How to Discount Double Counting When It Counts," *British Journal for the Philosophy of Science*, 59: 857–79.

Musgrave, A. (1974), "Logical Versus Historical Theories of Confirmation," *British Journal for the Philosophy of Science*, 25: 1–23.

Musgrave, A. (1989), "Deductive Heuristics," pp. 15–32 in K. Gavroglu, Y. Goudaroulis, and P. Nicolacopoulos (eds.), *Imre Lakatos and Theories of Scientific Change*, Kluwer, Dordrecht.

Worrall, J. (1989), "Fresnel, Poisson, and the White Spot: The Role of Successful Prediction in the Acceptance of Scientific Theories," pp. 135–57 in D. Gooding, T. Pinch and S. Schaffer (eds.), The Uses of Experiment: Studies in the Natural Sciences, Cambridge University Press, Cambridge.

Related Exchanges

Musgrave, A.D. (2006), "Responses," pp. 301–4 in C. Cheyne and J. Worrall (eds.), *Rationality and Reality: Conversations with Alan Musgrave*, Kluwer Studies in the History and Philosophy of Science, Springer, Dordrecht, The Netherlands.

Worrall, J. (2002), "New Evidence for Old," in P. Gardenførs, J. Wolenski, and K. Kijania-Placek (eds.), *In the Scope of Logic, Methodology and Philosophy of Science* (vol. 1 of the 11th International Congress of Logic, Methodology, and Philosophy of Science, Cracow, August 1999), Kluwer, Dordrecht.

Worrall, J. (2006), "History and Theory-Confirmation," pp. 31–61 in J. Worrall and C. Cheyne (eds.), *Rationality and Reality: Conversations with Alan Musgrave*, Springer, Dordrecht, The Netherlands.