Toward Progressive Critical Rationalism
Exchanges with Alan Musgrave

Deborah G. Mayo

Reacting to Alan Musgrave's challenges to my account of theories directs us to a distinct set of issues under the same broad headings as in my exchange with Chalmers in Chapter 2: how do we characterize and justify inferences beyond the data (ampliative or inductive inference), and how do we view the role of theory appraisal in characterizing scientific progress? Under the first, I wish to tackle head-on the problem of inductive or ampliative inference retained from Popper and other logical empiricists. Under the second, I explore the possibility that Musgrave's move toward a piecemeal account of explanation moves the Popperian critical rationalist to a more progressive standpoint. Hovering in the background is the metaphilosophical query: How do Musgrave's assumptions about inductive justification lead to his skepticism about induction?

1. *Experimental Reasoning and Reliability*: Can inductive or ampliative inference be warranted? Can we get beyond inductive skepticism by showing the existence of reliable test rules?
2. *Objectivity and Rationality*: Must scientific progress and rationality be framed in terms of large-scale theory change? Does a piecemeal account of explanation entail a piecemeal account of testing?
3. *Metaphilosophical Themes*: How do philosophical assumptions about the nature of reliable inference lead to skepticism about induction?

1 Brief Overview

I begin with a useful passage from Musgrave's contribution:

> So, the basic idea of critical rationalism is that it is reasonable to believe something if you have tried and failed to criticise it and show it to be false. . . . [This view] clearly

need[s] to be fleshed out with a theory of criticism – an account of the ways we can criticise explanatory hypotheses of various kinds. One way to criticise a hypothesis is to subject it to observational or experimental tests. Quite generally, criticism must be genuine or serious criticism, which could reveal that the hypothesis is false. So when it comes to empirical testing, the tests must be severe tests. Critical rationalism owes us a theory about what a serious or severe empirical test is. (Musgrave, pp. 104–5)

This is clearly very much in the spirit of my severe testing account. Note, however, that for a hypothesis *H* to have passed a "serious or severe empirical test" with test result *E*, it does not suffice that the test "could reveal that the hypothesis is false" by resulting in not-*E*. After all, this condition would be satisfied whenever *H* entails (or fits) *E*, and the test merely checks whether *E* results. Although such a weak falsificationist requirement is countenanced in a simple "hypothetical deductive" account, the key point of requiring severity is to demand more than that *H* entails *E* and that *E* occurs. If the test would very probably result in such a passing condition, even if *H* is false, then the test is insevere (by the minimal or weak severity requirement). Given that I view my account of testing as implementing and improving upon the Popperian idea of critical rationalism by cashing out the notion of a severe test, I have long been puzzled (and frustrated!) to find myself the object of Alan Musgrave's criticisms. As with all the essays on our exchanges, I focus on the arguments that have moved us forward, as well as those that highlight the problems that remain.

Although my focus is on Musgrave's discussion of severe tests, in fact it was his discussion of explanation in his contribution to this volume that finally allowed us to make real progress, as much as a year after the ERROR 06 conference. If one is prepared, as Musgrave is, to talk of inferring only parts of explanations, I reasoned, why stick to holistically inferring an entire theory rather than the portions that have passed stringent tests? And once the door is open to piecemeal testing in probing large-scale theories, much of the apparent disagreement evaporates. The final pages of Musgrave's contribution records the concordance that eventually grew between us, as regards accepting only certain pieces of a proposed explanation. A key question that remains *is whether his position on explanation can sit consistently with the comparativist testing account that he retains.*

Before getting to that, Musgrave raises key criticisms that usefully illuminate central issues. Some of Musgrave's criticisms were already taken up in my exchange with Chalmers, and, although I will not rehearse them, I will try to tease out their origins. Doing so reveals the deepest problem that has stood in the way of progress, not just by Popperians but by philosophers

of science more generally; namely, the problem of showing the existence of reliable or severe test procedures. This is a contemporary version of the problem of induction, which we may understand as the problem of warranting ampliative inferences – ones that generalize or go beyond the data.

Alongside these fundamental issues are matters that seem to me largely to reflect equivocations on terms; but not having convinced Musgrave that a disambiguation of meaning scotches his concerns, it will be important to take up some of the main ones in the sections that follow.

2 The Aims and Goals of Science

One of the chief motivations to turn to the new experimentalism and to scientific practice is to uncover the real-life driving forces of specific inquiries. I wrote, for example:

> The growth of knowledge has not to do with replacing or confirming or probabilifying or "rationally accepting" large-scale theories but with testing specific hypotheses in such a way that there is a good chance of learning something – whatever theory it winds up as part of. (This volume, p. 28)

Musgrave has a lot of fun construing my "not to do" as "nothing":

> Mayo thinks that . . . if a local and specific claim passes a severe test, this teaches us nothing whatsoever about the large-scale theory of which that claim might be a part. . . . Think what an extraordinary statement that is. (Musgrave, pp. 105–6)

My statement, however, was that the growth of knowledge "is not a matter of" assigning probabilities to, or rationally accepting, large-scale theories. Moreover, philosophical accounts that take those forms serve *at most* as rational reconstructions of science and miss what is most interesting and important about scientific progress, especially at the cutting edge. For one thing, we have to gather up a lot of experimental knowledge to even arrive at relevant data and reasonably probative tests. For another, we may not even have a large-scale theory when we set out – and often enough succeed – in learning about the world. In my view, "In much of day-to-day scientific practice, and in the startling new discoveries we read about, scientists are just trying to find things out" (Mayo, 1996, p. 56). Finding things out, moreover, includes finding out how to improve our capacities to probe, model, and detect errors and to make reliable inferences despite lacking a well-tested full-blown theory.

But, contrary to Musgrave's allegation, I hold that we learn a lot about the theory and about the domain in question even when only piecemeal hypotheses are passed severely. It is for this reason that I distinguished testing a large-scale theory from learning about it (Mayo, 1996, pp. 189–92). In the case of experimental general relativity, for instance, scientists had to develop the parameterized post-Newtonian (PPN) framework, estimate lots of parameters in statistical models, and learn about instruments and about effects that must be accounted for, subtracted, or dismissed later. This is theoretical knowledge all right, but it does not amount to accepting an entire theory or background paradigm or the like. Quite the contrary, the goal is to set out a testing framework that does not require assuming the truth of any theories not yet well corroborated.

A supposition that seems to underlie Musgrave's, and perhaps also Chalmers's position, is that anyone who does not set out to build an account of theory acceptance must be an antirealist, viewing the aim of science as limited to "empirical adequacy" perhaps along the lines of van Fraassen (1980). That is why Musgrave claims I am a critical rationalist at the observable level and wonders how I can avoid moving up to the theoretical level – but I never made any such distinction. The models that link data to hypotheses and theories all contain a combination of observational and theoretical quantities and factors, and I am glad to move "up" insofar as stringent testing allows. The entire issue of scientific realism looms large in Musgrave's account; it is one about which the error statistician may remain agnostic, and, thus, I leave it open.

3 Justifying *H* versus Warranting Test Rules That Pass *H*

Among the most puzzling of Musgrave's statements has to be this:

> My question is – can we incorporate [Mayo's] theory into critical rationalism? Deborah Mayo does not think so. She rejects critical rationalism. She rejects it because she accepts justificationism. The upshot is that she abandons the project of appraising large-scale theories altogether. I want to entice her back into the critical rationalist camp. (Musgrave, p. 105)

"Justificationism," as Musgrave uses that term, holds that to warrant a hypothesis requires either showing that it is true (error free) or that it is probable, in some sense which he does not specify. Musgrave seems to think that my denial of comparativist testing of theories is explicable only by my being a justificationist! But I too reject justificationism as he defines it. A hypothesis is warranted for an error statistician not by being justified as

highly probable but rather by having been highly probed, that is, by passing a stringent test (Mayo, 2005). The basis for rejecting the comparativist account of theory testing, as argued at length in Chapter 1, is that to infer a theory simply because it is the best tested thus far is a highly unreliable rule. In other words, if we go along with Musgrave's use of "justifying *H*," then we agree that the task at hand is not to justify *H* but rather to warrant a rule that outputs hypothesis *H*. The difference between us – and it is a significant one – is that, for me, a warranted test rule must be reliable (i.e., it must be a reliable probe of relevant errors).

Once it is understood that to warrant the test that outputs *H* requires the test rule to be reliable or severe, the critical rationalists' insistence on banishing talk of justifying hypotheses appears silly. A hypothesis may be said to be justified (in our sense) by having passed a warranted (severe) test rule. The trouble for critical rationalists is that they deny it is possible to warrant a test rule as reliable; for that would involve an inductive justification, which they do not allow themselves. In keeping with this standpoint, Musgrave feels himself bound to reject the need for any kind of inductive or ampliative claim even at the "meta-level," as he calls it – that is, even as regards the reliability of test rules or methods of criticism. Yet showing meta-level reliability would seem to be at the very foundation of implementing critical rationalism.

3.1 Progressive Critical Rationalism

I wish to entice Musgrave to advance the critical rationalist program by actually developing rules for good and severe tests and for distinguishing them from illicit, unreliable, and unscientific test rules. Musgrave himself claims that "critical rationalism owes us a theory about what a serious or severe empirical test is." (p. 105) The time is ripe to provide it! To avoid confusion, we need to distinguish what critical rationalism has been since Popper – *Popperian critical rationalism* – and a forward-looking theory of criticism, which we may call *progressive critical rationalism.* (Note: I leave to one side the more radical Popperianism espoused by David Miller but soundly rejected by other followers of Popper.) Progressive critical rationalism would proceed by developing tools for severe tests. Such tools seek reliable probes of errors, and yet, in another verbal misconstrual, Musgrave imagines that I mean "proved" (with certainty) when I say "probed"!

But if a theory does not fail a severe test, Mayo does not want to say that the whole theory has passed the test. Why not? Because Mayo thinks "passing a severe test"

means "certified by that test to be correct." She wants to know what passing a severe test proves to be true. Passing a severe test does not prove the theory tested to be true. Therefore, the theory tested has not passed the test. (Musgrave, p. 105)

To this indictment, I am left to repeat the same baffled cry I let out in considering Chalmers. I would not be appealing to statistical inference were I to think that *H*'s passing a severe test "proves" *H*. It suffices that *H* passes despite being put to stringent probes of the ways in which *H* may be false. If, after a series of stringent probes and deliberately varied results, hypothesis *H* holds up, then we may say *H* has passed severely – this entitles us to infer *H*, even though we qualify the inference by means of the severity assessment of the test rule (formally or informally).[1] In some cases, the inferred claim may assert how far off *H* is from a correct description of the process in question. With Musgrave's interest in talking of approximate truth, this would seem just the ticket.

I am happy to use Popper's more succinct "*H* is corroborated" interchangeably with "*H* has passed a severe test" so long as my notion of severity is understood.[2] Whole theories may be corroborated; my objection is only to regarding a theory as having passed severely when it is known that it makes claims that have been probed poorly or not probed at all. Now Musgrave makes clear, here and in an earlier exchange, that critical rationalism "proposes that it is reasonable to believe (adopt, prefer) that theory, if there is one, whose consequences have been best tested" (Musgrave, 2006, p. 307). If one adopts this rule for "reasonable believings" (as Musgrave sometimes refers to them), it would be reasonable to believe all of the General Theory of Relativity (GTR) in 1970, say, because its consequences about the deflection of light had held up to testing. Such a rule for "reasonable believings" would be highly unreliable.

Is Musgrave's critical rationalist not bothered by the fact that he advocates a rule for inferring or believing that is known to be unreliable? In Musgrave (1999, p. 346) he said No, but in his (2006) response he agrees that he would be bothered. He now clarifies his position as follows: Although "critical rationalists need not assert or prove that their method is reliable before they can rationally adopt it . . . if the method can be criticized by showing

[1] Note an error in the first (and last) sentence of his quote above, although perhaps this is just a slip. We distinguish between passing a test and passing it severely. If the test is severe for all of *H* – meaning that *H* as a whole has passed severely with the test and data in question – then all of *H* has passed severely. Yet even if the test is insevere, I am prepared to say that *H* has passed, just inseverely.

[2] Popper at one point allowed that one was free to call an inference to a severely tested hypothesis *H* an inductive inference to *H*, although he did not do so himself. C.S. Peirce talked this way long before.

it to be *un*reliable, then they should cease to employ it" (p. 308). This is an important point of progress. However, it only spells trouble for Musgrave's critical rationalist because it would indict the comparativist rule of testing as he defines it.

Musgrave could solve this problem by replacing "best tested" with severely tested in my sense. He would then have advanced to what I am calling progressive critical rationalism. In the wrap-up to this essay I consider whether he has moved at least partway to this standpoint. For now I need to continue with Musgrave's fairly strenuous resistance to buying into this account.

3.2 Can Inductive Generalizations Be Warranted?

The elephant in the room is the fear of endorsing any kind of "induction." Because of this fear, Musgrave seems ready to claim that no test based on finitely available evidence can ever severely pass a claim that goes beyond that finite evidence. It is the same stance that we find with Chalmers, whom Musgrave quotes, approvingly, as having pointed out that data collected in Baltimore only warrants with severity assertions about how well the claim held in Baltimore: "Suppose I check some experimental generalization in my laboratory in Baltimore on a Tuesday afternoon, and it passes the test." According to Musgrave, this passing result cannot warrant with severity the general claim because "it has implications, hitherto untested, about what will happen if I do the same test on Wednesday afternoon, or in Buenos Aires instead of Baltimore." (Musgrave, pp. 108–9)

Think what an extraordinary statement that is. The chemical shown to be toxic in the laboratory in Baltimore only teaches us what happened in Baltimore? Notice, one could strictly only "infer" the results already observed! This is vintage Popper. Popper made it clear that for him corroboration was merely a report of past success; there was no warrant for ever *relying* on claims thereby passed (since that would allude to trials other than those run). Although this may seem surprising, in fact this is the skeptical standpoint that the critical rationalist recommends. Rather than reconsider Popper's scepticism, critical rationalists seek notions of scientific rationality that can live alongside this scepticism.

3.3 The "Wedge" between Skepticism and Irrationalism

This takes us to the distinction Musgrave draws between warranting a rule (or a procedure) for "believings" or "inferrings" as opposed to warranting the claims believed or inferred. Musgrave regards this distinction as the

secret by which Popper is able to avoid the problem of induction. According to Musgrave, "The way out of this impasse is to adopt critical rationalism and reject justificationism. What needs 'warranting,' if we must use that trendy term, is not the theories but our adoption of them." In this way, Musgrave allows, we can "warrant believing the theory, though not the theory itself" (p. 107). Many will wonder what it can mean to say:

We are warranted in believing hypothesis H, although belief in H is not warranted.

If someone says "I am warranted in believing that X is toxic to the liver" while denying that he regards the assertion "X is toxic to the liver" to be warranted, it would seem he is playing a verbal game. Yet, this is the kind of convoluted language Popperian critical rationalists seem forced to adopt because they assume warranting H would mean justifying H as probably true. Therefore, the way to parse the above sentence is to replace the second "warranted" with "justified" in the justificationist sense. Then the claim is as follows:

We can warrant a rule for accepting (or believing) H without claiming that H is proved true or probable.

The assertion now is perfectly reasonable; moreover, it is one the error statistician happily endorses. Error statistics, after all, is based on adapting the idea from Neyman-Pearson (N-P) testing that what we want are rules for "deciding" to classify hypotheses as "accepted or rejected," and probabilistic properties attach to the rules, not to the hypotheses inferred. For instance, in 7.1 of the introductory chapter, we considered a test rule T that maps data into the claim "there is evidence for a discrepancy from μ_0." Where we differ from Musgrave concerns what is required to warrant a test rule (or, as he might prefer, a procedure for believing). We require the rule be sufficiently reliable or stringent or the like; whereas the Popperian critical rationalist deems this goal beyond reach. Fittingly, Lakatos once said that N-P tests are excellent examples of Popperian "methodological falsificationism" (Lakatos, 1978, p. 109, note 6). Unfortunately, the Popperians never made use of the fact that N-P tests, in the realm of statistics, provide rules with low probabilities of erroneously rejecting or erroneously accepting hypotheses. If they had, they might have seen how to adapt those statistical rules to their purposes. That is essentially what I have tried to do.

A reformulation of the strict (behavioristic) construal of N-P tests is required because tests may have good long-run error probabilities while doing a poor job of ruling out errors in the case at hand (see, e.g., Chapter 7, this volume). Still, the formal determination of test rules with good error

probabilities is extremely important because it demonstrates that stringent or reliable test rules exist, which lets us check putative rules for reliability. Numerous examples of how error properties of test rules relate to the severity of particular inferences are discussed throughout this volume.

To make some quick points, consider a formal test rule R that rejects a null hypothesis:

H_0: drug X is *not* toxic (e.g., to rats),

just when a randomized controlled trial yields an increase, $d(\mathbf{x}_0)$, in risk rates that is statistically significant at the .01 level. In those cases, $d(\mathbf{x}_0)$ is taken as grounds for rejecting H_0 and inferring H^*, where

H^*: there is evidence of the increased risk.

Then H^* has passed a severe test because

Prob(rule R would not infer H^*, when in fact H_0 is true) $= .99$.

Musgrave's random sample of rats in Baltimore is effectively a random sample of all rats with respect to this risk; the error probabilities apply for subsequent applications so long as the statistical model assumptions hold sufficiently, and this is testable. Now someone may charge that, although the generalization about X's toxicity holds for all rats today, X may not be harmful at some future date – but this is not problematic for our account. It is precisely because we can vouchsafe that the test rule continues to operate in a stable way that we can discover the effect has changed!

It is not clear why Musgrave charges that experiments on GTR fail to warrant general deflection hypotheses. I chose to discuss such "null tests" in physics precisely to apply the error-statistical account to the home turf of the high-level theory philosophers. Here the null hypothesis is set as a precise claim – one from which it would be very easy to detect departures:

H_0: deflection effect $= 1.75$(GTR prediction)

Having observed $d(\mathbf{x}_0)$ that accords with H_0, a test rule R infers that

H^*: the deflection effect differs no more from the GTR value than $\pm\varepsilon$,

where, say, ε is three standard deviations away from $d(\mathbf{x}_0)$. If one wishes to speak only in terms of falsifications, there is no problem in doing so: inferring or corroborating H^* is the same as falsifying the denial of H^*. We have

Prob(rule R corroborates H^*, when H^* is false) $= .001$.

Therefore, when R corroborates H^*, H^* passes the test with severity .999 (see also Introduction, p. 22, and Chapter 7, this volume). Thus, our warrant of the reliability of rule R is no different than warranting the rule for rejecting the zero-risk-increase hypothesis. Unless Popperians wish to discard the logic of falsification, they cannot object to corroborating claims based on falsifying their denials!

Popperian critical rationalists make the success of science a mystery (indeed, Popper calls it "miraculous"). It is no great mystery that statistical (and other) predictions hold in particular cases of a type of experiment because if we had failed to bring about the experimental conditions for testing the prediction, then we would discover our errors, at least with high probability. Scientists build their repertoires for checking instruments and models so that errors not detected by one tool will be picked up by another.

3.4 Beyond Formal Inductive Logics

Popper spoke of corroborating claims that passed sincere attempts to falsify them, but he was at a loss to operationalize "sincere attempts to falsify." Merely "trying" to falsify, even if we could measure effort, would be completely beside the point: one must deliver the goods. It is the properties of tools that have to be demonstrated (i.e., that were a specified discrepancy present, it would have been detected with high probability). Combining different tests strengthens and fortifies the needed claims.

One can perhaps understand why Popper remained skeptical of identifying reliable rules: at the time, philosophers thought that one needed a formal inductive logic that would be context-free, a priori and not empirical. Pursuing the path of formal inductive logics failed to yield applicable, useful rules, so the whole project of developing reliable rules and methods was abandoned. The time is ripe to get beyond this. We invite Musgrave and other critical rationalists to turn to the project of identifying reliable rules of criticism!

4 A Constructive Ending: Explanation and Testing

By the end of his chapter, Musgrave seems prepared to adopt what I call progressive critical rationalism, if only because it emerges naturally out of his discussion of inference to the best explanation (IBE) in earlier sections. For one thing, in his discussion of IBE, it is required that the predictive success to be explained by H be surprising or novel: First, if it is very easy

to achieve explanatory success H would not pass severely. Second, when it comes to inferring explanations, Musgrave is prepared to infer pieces or aspects of theories. Here inference to parts is not only allowed but is welcomed. Indeed, Musgrave recommends replacing Popper's search for a measure of verisimilitude, understood as closeness to the truth, with partial truth understood as truth of the parts:

> Just as total success is best explained in terms of truth, partial success is best explained in terms of partial truth. . . . That true part, common to both theories, sets out the relative motions of Earth, Sun and superior planets. (Musgrave, p. 101)

The position is clearly in sync with our piecemeal theory of severe testing: we may infer those aspects or variants or subparts of a theory that hold up to stringent scrutiny.

In his last section Musgrave hints at renouncing his advocacy of inferring the entire theory when only parts have been severely corroborated. He considers the addition to a hypothesis T of another hypothesis X that broadens or extends the domain of $T - X$ is not what he calls an "idle wheel" or irrelevant conjunct. (The "tacking paradox" returns in Chapter 9.)

> Of course, if X is not an "idle wheel," and if $T\&X$ has regions of implication that T does not, which have not yet been checked, then critical rationalists will not advocate acceptance of $T\&X$ over T. As Mayo rightly insists, they require evidence that these excess implications are correct. (Musgrave, p. 110)

Musgrave seems to be leaving off with the promise that he is ready to make the move to progressive critical rationalism, thereby attesting to the fruitfulness of multiple exchanges!

References

Lakatos, I. (1978), *The Methodology of Scientific Research Programmes*, J. Worrall and G. Currie (eds.), Vol. 1 of *Philosophical Papers*, Cambridge University Press, Cambridge.

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

Mayo, D.G. (2005), "Evidence as Passing Severe Tests: Highly Probed vs. Highly Proved," pp. 95–127 in P. Achinstein (ed.), *Scientific Evidence*, Johns Hopkins University Press, Baltimore.

Mayo, D.G. (2006), "Critical Rationalism and Its Failure to Withstand Critical Scrutiny," pp. 63–99 in C. Cheyne and J. Worrall (eds.), *Rationality and Reality: Conversations with Alan Musgrave*, Kluwer Studies in the History and Philosophy of Science, Springer, The Netherlands.

Mayo, D.G. and D.R. Cox (2006), "Frequentist Statistics as a Theory of Inductive Inference," pp. 77–97 in J. Rojo (ed.), *Optimality: The Second Erich L. Lehmann Symposium*, Lecture Notes-Monograph Series, Institute of Mathematical Statistics (IMS), Vol. 49.

Musgrave, A. (1999), *Essays in Realism and Rationalism*, Rodopi, Amsterdam.

Musgrave, A. (2006), "Responses," pp. 293–333 in C. Cheyne and J. Worrall (eds.), *Rationality and Reality: Conversations with Alan Musgrave*, Kluwer Studies in the History and Philosophy of Science, Springer, The Netherlands.

van Fraassen, B. (1980), *The Scientific Image*, Clarendon Press, Oxford.

Related Exchanges

Chalmers, A. (2006), "Why Alan Musgrave Should Become an Essentialist," pp. 165–82 in C. Cheyne and J. Worrall (eds.), *Rationality and Reality: Conversations with Alan Musgrave*, Kluwer Studies in the History and Philosophy of Science, Springer, The Netherlands.

Worrall, J. (2006), "Theory-Confirmation and History," pp. 31–62 in C. Cheyne and J. Worrall (eds.), *Rationality and Reality: Conversations with Alan Musgrave*, Kluwer Studies in the History and Philosophy of Science, Springer, The Netherlands.