Science Is Irrational—and a Good Thing, Too

Michael Strevens

Forthcoming in *Extreme Philosophy*,
edited by Stephen Hetherington, Routledge

Abstract

Many scientists tout the probative role of theoretical beauty—that a theory is beautiful, they say, gives us considerably more reason than we would otherwise have to believe it. At the same time, appeals to beauty are systematically excluded from the journals and conference proceedings that function as the official organs of scientific discourse. Such exclusions, along with a number of other kinds of censorship, violate the principle of total evidence. Perversely, this departure from rationality plays an essential role in undergirding science's powers of discovery, by focusing researchers' efforts as intensely as possible on the production of empirical data, to science's long-run benefit. After presenting this case for the strategic irrationality of the rules of scientific publication, I rebut attempts to argue that, in view of their functionality, science's breaches of the total evidence principle are in fact rationally permissible.

1. Truth and Beauty

Among Nobel Prize–winning theoretical physicists, testaments to the probative power of beauty abound:

> It is more important to have beauty in one's equations than to have them fit experiment (Paul Dirac)
>
> [Beauty is] a chief criterion for the selection of a correct hypothesis (Murray Gell-Mann)

> We would not accept any theory as final unless it were beautiful (Steven Weinberg)[1]

Some, indeed, have penned entire books in praise of beauty's power to guide us to truth (Chandrasekhar 1987; Wilczek 2015). There is a widespread belief, then, that all other things being equal, and in physics at least, a more beautiful theory—in perhaps some very particular sense of "beautiful"—is more likely to be true.[2]

Further, physicists give more than lip service to beauty; they do in fact attend to it in their deliberations, as explained by another theoretical physicist, Brian Greene:

> [Physicists] make choices and exercise judgments about the research direction in which to take [a] partially completed theory [that are] founded upon an aesthetic sense—a sense of which theories have an elegance and beauty of structure on par with the world we experience (Greene 1999, 166–7).

At the same time, as Greene goes on to say, such investigative techniques are sidelined in scientific argument:

> Aesthetic judgments do not arbitrate scientific discourse. Ultimately, theories are judged by how they fare when faced with … hard experimental facts (p. 166).

That all seems a little contradictory. How can it be that physicists are guided by aesthetic concerns—evaluating their ideas in part by criteria that advert to beauty, elegance, and all that—and yet at the same time, that their theories are not even in part judged by such concerns? Don't the acceptance of such guidance and the making of such judgments amount to more or less the same thing?

1. Respectively: Dirac (1963, 47); Johnson (1999, 239); Weinberg (1992, 165).
2. There are dissenters, such as Sabine Hossenfelder (2018).

To understand the situation, we need to distinguish two different venues for scientific deliberation. On the one hand, there is a scientist's own private reasoning and informal discussions with other scientists. On the other hand, there are science's official organs of communication, the journals and conference proceedings. The rules governing these two venues are rather different. In private reasoning, aesthetic evaluation is legitimate, and as Greene observes, rather common. In official channels, aesthetic evaluation is, quite simply, banned—which is to say that it is impermissible, in a journal article laying out a hypothesis and the reasons to accept or reject it, to appeal to considerations of beauty in any way. Empirical success alone may be cited. In the pages of the journals, then, the evaluation of a theory must focus exclusively on "hard experimental (or observational) facts". This is what I have elsewhere called science's *iron rule*, the stricture that in official scientific argument, only empirical testing counts (Strevens 2020).

As Greene's words show, theoretical physicists are quite accustomed to this state of affairs. In their own thinking they follow the aesthetic signposts, but then they argue for their theories on entirely empirical grounds in the journals.

Yet there is something more than a little awkward about this convention: it denies physicists such as Murray Gell-Mann, Steven Weinberg and so on the right to invoke, when seeking to convince others of the soundness of their ideas, the aesthetic cues that they themselves find so persuasive. Their deepest reasons for belief are to be expunged from their official advocacy. This situation is, indeed, not merely awkward but literally irrational—specifically, a violation of what Carnap called the principle of total evidence.

The total evidence principle enjoins us to take into account, when deliberating, all relevant evidence. Or in a more humanized version, it requires us to take into account all relevant evidence that is not too expensive to obtain or put to use, relative to the importance of the issue to be decided. In short: do not ignore any accessible source of pertinent evidence.

Judgments about theoretical beauty and elegance are surely inexpensive: a scientist who understands a theory need do little in addition to evaluate its aesthetic merits, and these merits are then relatively easy to incorporate into deliberation—they require very little in the way of storage or computational power. To ignore them, then, is to throw away information that can be had cheaply and yet which is thought by many scientists to be of great value. That is what science's iron rule tells us to do; that is against the canons of reason.

It is the purpose of this paper to defend the claim that the iron rule is irrational. Let me make a few remarks to clarify the scope and nature of the irrationality.

First, whether scientists' trust in beauty is justified is not important to my thesis. I see the principle of total evidence as having an internalist character: you should take into account all (important, not prohibitively expensive) evidence that *you regard* as relevant.

Second, it is unimportant whether this quality of theories that scientists call "beauty" is genuinely a kind of beauty. Perhaps it is in many cases a kind of symmetry that is, strictly speaking, aesthetically neutral. What makes for the iron rule's irrationality is that this quality is believed to be probative, and yet it is interdicted by the rule.

Third, I emphasize once again that the iron rule does not forbid the scientist to attend to aesthetic concerns in their own reasoning. It applies only to what they write in the official record—in the journals and so on. Thus, it imposes irrationality only on the canons of public scientific argument.

That brings me to the question of the locus of this irrationality. I believe that science is a reliable form of inquiry, and indeed, that its reliability is in part due to its imposition of the iron rule. Science is consequently not an irrational enterprise: those who would seek to understand the workings of the universe are well advised to become scientists. Further, in so doing, they are not thinking or acting irrationally. Science's irrationality subsists not in the enterprise as a whole, then, nor in the thoughts or actions of its practitioners,

but in its rules of public engagement—in the procedures it lays down for the official, public presentation of reasons to accept or reject its many models, hypotheses, and theories.

2. The Radical Incompleteness of Official Scientific Argument

The beauty ban is but one facet of the irrationality of official scientific argument. There are many other ways in which official argument—the presentation of considerations for and against hypotheses in the journals and conference proceedings—is unreasonably restrictive, excluding relevant and readily available information and therefore violating the total evidence principle.

These irrational exclusions stem from two aspects of the iron rule. The first is the rule's exclusion of all superempirical virtues from official argument. Beauty is (or is considered to be) one such virtue; others include certain kinds of explanatory unity and parsimony. The iron rule simply forbids the invocation of any such considerations in official argument. Yet unity and parsimony are, like beauty, considered by many scientists to be useful to the evaluation of theories (Schindler 2022).

This point deserves a more subtle discussion that I can give it here, as there are many appeals to explanatory unity and simplicity that are empirical in character, and that are therefore permitted by the iron rule. Some invocations of parsimony, for example—such as those inherent in statistical measures such as the Akaike Information Criterion—are thought to be justified because they help scientists to avoid overfitting their hypotheses to the data. Fit is an empirical issue, bearing as it does on the degree to which observations support various hypotheses. An adequate treatment of the iron rule's ban on superempirical arguments, then, must carefully unweave the empirical from the superempirical in appeals to simplicity, unity, and so on. (Even the case of beauty is perhaps not entirely straightforward.) That is not the business of the present paper, however, so having identified the broad contours of the ban, I will move on.

The second way in which the iron rule effects an unreasonable narrowing of scientific argument is by imposing a desideratum that scientific writing be as objective as possible, in the sense that it ought to introduce only considerations that scientists can in principle agree on at the time of publication. Included under this rubric are three rather different kinds of statement: assertions of states of affairs that are widely accepted by the scientific community; assertions about what is observed that depend minimally on disputable background assumptions, such as specifications of measurements made with an experimental apparatus, the conditions obtaining when those measurements were made, and so on; and finally, statements about formal or mathematical measures, such as degrees of statistical significance. Other matters, stipulates the iron rule, should be avoided as far as possible. (Think of "objectivity" as a term of art, not as denoting a property of independent philosophical interest.)

The boundaries of the objective in this sense are of course rather loosely defined; it is up to peer reviewers and other gatekeepers of the official scientific literature to enforce them as best they can. As with the superempirical virtues, I won't aim, in this paper, for any more detailed or precise statement of what the desideratum of objectivity amounts to; rather, I want to examine a certain consequence of the desideratum: the hollowing out of scientific argument. Let me illustrate with an example.

In the wake of the First World War, the British physicist Arthur Eddington led an expedition to test Einstein's new general theory of relativity by photographing stars close to the sun to determine the degree to which light rays were deflected by gravitation. Such measurements could be made only if the sun were blotted out—for which reason Eddington and his collaborators set out with their telescopes and cameras to make their measurements during the total eclipse of the sun that would occur in parts of South America and Africa in May 1919.[3]

3. Stanley (2003) gives an illuminating account of the background and conduct of the experiment; Earman and Glymour (1980) discuss the weaknesses of the experiment; Strevens

While Eddington himself set up his apparatus on the African island of Príncipe, other members of his team took two telescopes to Sobral, in north-eastern Brazil. With these instruments, the Brazilian team obtained two separate sets of photos of the star field behind the sun during the few minutes of totality. The results were at odds: one telescope, the "astrographic", suggested a bending angle for light that was almost exactly in accordance with the predictions of Newtonian physics (as Eddington extrapolated them), while the other, the "4-inch" suggested a bending angle more in accordance with the predictions of Einstein's theory.

The two sets of photos were not, however, of equal quality. Something had gone wrong with the astrographic setup: the images were somewhat blurred. In their official presentation, the experimenters attributed the fuzziness of the photos to a distortion of the telescope's mirror caused by the heat of the sun, and then continued: "it is difficult to say whether this caused a real change of scale in the resulting photographs or merely blurred the images" (Dyson et al. 1920, 309). That difficulty is a significant one: if there were nothing more than a blurring due to a slight loss of focus, the plates would still have contained valuable information about the apparent positions of the stars and thus about the bending power of gravity, but if there were a "change of scale", that is, a systematic displacement of the stars' positions on the plate, the results would be more or less useless.

Upon this hard-to-decide question, then, rested the distinction between a mixed result in which one instrument aligned with Einstein and the other with Newton, and a result that clearly favored Einstein. Eddington and his collaborators opted decisively for the latter interpretation, declaring in their final paragraph that their measurements left "little doubt" that Einstein's prediction was correct. That conclusion makes sense only if there is equally "little doubt" that the blurring of the astrographic plates was caused by a change

(2020) points specifically to the gap in the argument to be discussed below. Collins and Pinch (2012) present an influential but somewhat uncharitable view of the experiment.

of scale.

How did the authors reach such a verdict? They don't say. Some writers have supposed, in retrospect, that they indulged in motivated reasoning, as Eddington is well known to have favored Einstein's theory from the very start. For the sake of the argument, however, let us suppose they had their reasons. Why do they not appear in the paper?

The answer, I believe, is the iron rule's curb on subjectivity. The authors present what objective information they can: the fact of the blurriness of the plates, and the uncontroversially plausible suggestion that the heat was a likely cause of the blurriness. They had further information, I assume, that inclined them to suspect that the heat caused a change of scale. But that information did not take the form of observable facts, but rather of certain subjective estimates—it was, if you like, a subjective probability distribution conditioned on various observations of the properties of the setup and the consequences. As such, it could not be reported—it would be a blatant violation of the iron rule's injunction to remain objective. What cannot be reported must be passed over in silence. Thus, there is a curious lacuna in the argument of the Eddington paper, at a most sensitive and crucial point.

This is far from an idiosyncrasy. Scientific papers almost never contain all the information relevant to assessing the weight of the reported evidence. As in the case of beauty, then, readers are systematically denied access to considerations that the authors consider to be of the greatest importance for evaluating the theories under examination—a flagrant violation of the principle of total evidence. Is there some rationale for this—some way in which the iron rule, by suppressing apparently valuable information, contributes to science's truth-finding power? Yes, there is.

3. The Benefits of Irrational Narrowness

The greatest of Thomas Kuhn's insights into the nature of science was that the institutional rules of science "force scientists to investigate some part of nature

in a detail and depth that would otherwise be unimaginable" (Kuhn 1996, 25). It is because of this detail and depth, Kuhn believed, that scientists are able to turn up the little "anomalies" that eventually precipitate scientific revolutions and propel science onward through a sequence of ever more accomplished theories.

The institutional rules that Kuhn has in mind are of course those prescribed by the "paradigm", a practical and theoretical framework that turns any area of science into a collection of "puzzles" to be solved. For a scientist committed to a paradigm, both the importance and the solvability of these puzzles are taken for granted. All that's left for the scientist to do is to select a puzzle and to throw everything they have into its solution. The paradigm's puzzle-posing framework thereby channels vast quantities of scientific energy into those many questions of subtle detail, among which lie the clues to the next great theoretical breakthrough.

Few historians, sociologists, or philosophers would now endorse the Kuhnian picture in all its aspects, and the puzzle-solving conception in particular has not survived scrutiny. Whereas the correctness of a proposed solution to a puzzle can be assessed quickly and without controversy, it is common for the results even the most "normal" scientific research to be doubted or at least regarded warily by some other scientists in the field.[4]

Yet I think that Kuhn put his finger on a point of great importance. There is something about the norms of science that focuses enormous energy on problems that are expensive, time-consuming, and often both tedious and risky to pursue, and that are in many cases not obviously "intrinsically interesting and important" (Kuhn 1996, 37)—problems that nevertheless occasionally turn out to be crucial for scientific progress.

That "something" is (and here I depart from Kuhn) the iron rule's insistence that scientific argument be framed, in its official form, objectively

4. Biagioli (2012) provides an overview of Kuhn's significance fifty years after the publication of *Structure*.

and with reference to empirical testing alone. Although this stricture applies only to scientific communication, and not to scientists' private thinking, it makes the public side of science into something of a game. The rules of the game are, because of the narrowness of the iron rule, rather constraining: a legitimate move—a published paper—must be concerned solely with contributing in some way to an empirical argument for or against a theory, where that argument must be framed as far as possible without subjective elements. (Contributing to an empirical argument need not mean making observations; it might mean framing theories that are to be empirically tested or developing measurement apparatus or statistical techniques to be used in such tests.)

Consequently, if scientists wish to play the game, they must direct their energies toward framing or carrying out empirical tests. That aspect of the iron rule which says that "only empirical testing counts", forbidding any appeal to beauty or to the other superempirical virtues, thus performs much the same function as one of Kuhn's puzzle-posing paradigms, funneling scientists' energy and attention into the minutiae of prediction and measurement.

Eddington, for example, was far more of a theorist than an experimenter. (A stint at the Cavendish Laboratory in his early career seems to have ended ignominiously.) He appreciated Einstein's theory for its mathematical beauty and its lovely explanation of the equivalence of inertial and gravitational mass. He wanted to convince the world of its truth. But he was playing the science game, so he needed to put aside his mathematical subtlety and aesthetic sensitivity to make the long journey to Príncipe, devoting many months of his life to humdrum preparations for a few moments of eclipse totality during which he could only hope that the skies were clear. (They were not. Fortunately, the Brazilian clouds were better behaved.)

The iron rule's insistence on objectivity expedites the funneling, by freeing scientists from the need to formulate and deploy arguments for contentious premises in their written reports. However important such premises may be to the logic of their case, they are suppressed. That relieves researchers of the

tremendous burden of managing the social complexity of doubt, persuasion, and dissent. (Philosophers in particular should appreciate just how much effort, agony, and risk is involved in such undertakings.) Their labor goes entirely into producing their data—where, as the history of science shows over and over again, it stands to make the greatest and the most decisive difference.

Those aspects of the iron rule that I have identified as irrationally narrow lie at the heart of this arrangement. Eddington had to get on the boat to Príncipe because the arguments from beauty and elegance, which he regarded as highly persuasive, were forbidden. And the ease with which his results could be published depended in great part on the gaping hole in his paper's argument: the desideratum of objectivity spared him from having to make a case that the astrographic telescope suffered a change of scale. His reasons for favoring Einstein's theory may have been censored, but in the end what changed physics was not his aesthetic argument or his team's hunches about what went wrong with the astrographic telescope. Revolution was precipitated by the measurements, and the measurements of similar experiments that came after. It is an intense and exclusive emphasis on empirical testing, not logical rigor, that makes modern science special.

4. From Functionality to Rationality?

If the logic-defying element of the iron rule is so fruitful, can it really be all-things-considered contrary to reason? I aim to answer in the affirmative: the iron rule is both enormously productive and, for all that, irrational.

Before I continue, let me remind you of the locus of the alleged irrationality. The scientific enterprise as a whole is not unreasonable; nor are its practitioners. Science makes sense as an epistemic technology, and those who make use of its techniques in the light of its success do so in full accord with the prevailing epistemic norms. What is I claim to be irrational, because it violates the principle of total evidence, is the set of guidelines orchestrating official scientific publication—the iron rule itself.

A humanized version of the total evidence principle—a version that ordinary, conscientious people can reasonably be expected to follow—will make exceptions in cases where the contemplation of relevant information is disproportionately costly. You walk into a restaurant, and the host tells you to sit anywhere you like. You might stand there for ten minutes, carefully considering the pros and cons of each table, but no one will consider you irrational if you satisfice, heading straight for a table that is clearly good enough. It is simply not worth the cost of computing which table is absolutely the best. (Plus, of course, there is the risk that your date might take you to be a tiresome neurotic.)

Can we apply the same epistemic charity to the iron rule? The line of reasoning might run as follows. The considerations excluded from official scientific argument by the iron rule are relevant to the matters at hand—that must be conceded. But to take them into account would nevertheless impede the advance of science: investigators would get caught up in the construction of arguments from aesthetics and explanatory unity, or in the more controversial aspects of the interpretation of the evidence, and so would be wastefully diverted from doing what is more efficacious, namely, redoubling their efforts at observation, measurement, and experiment. Like the vacillator in the restaurant, they may produce a more completely reasoned piece of writing in the short term, but at a cost that ultimately more than outweighs the benefits. A humanized total evidence principle will therefore exempt them from such counterproductive epistemic duties, and so free them to follow, with a clear epistemic conscience, the iron rule.

Such considerations surely do vindicate the institution of iron rule–governed science as a whole, alongside the individual scientist's decision to participate in the institution. Equally—indeed, what is more or less the same thing—they establish the reasonableness of practitioners' committing to obey the iron rule in all of their official writing, without exception.[5] It does not follow, however,

5. Or perhaps with exceptions only in the most extreme cases, for example, when their

that each individual enactment of the rule is itself rational.

To see what I mean, consider Eddington's predicament as he sits at his desk preparing his report on the eclipse experiment for publication. Suppose that he has some information ("subjective", in the sense that matters to the iron rule) that bears on the question whether the Brazilian astrographic telescope suffered a change of scale. It is at his fingertips; should he write it down? The opportunity cost for doing so is perhaps not so very large. It would take only a few minutes to pen those paragraphs. Or suppose that he has in hand a marvelous aesthetic argument for the general theory of relativity. Should that go into the paper? (Or perhaps into a companion piece?) It might take only a few days of his time to put it into words—a small fraction of the year or so he spent organizing and carrying out the eclipse experiment.

Any reasonable cost-benefit analysis will answer yes to both questions. The effort is small and the reward, in terms of a more complete and useful case in favor of general relativity, is surely much larger. So, the information proscribed by the iron rule ought to be (or at least, is permitted to be) added to the official record. To prevent its being aired is a violation of the total evidence principle; that the iron rule does so in this case is irrational.

How can this be if, as I have proposed, the construction of aesthetic and other superempirical arguments, and the management of the subjective side of inductive reasoning and evidential support in general, is in the long term more of a hindrance than a help? Why doesn't the humanized version of the total evidence principle take this inefficiency into account, as suggested above, excusing the apparent violations of the principle on the grounds that in the long run, they would be unreasonably expensive, in terms of other evidence foregone, to avoid?

If the rules of rationality operated along thoroughly consequentialist lines, taking into account all aspects of context within the longest possible time horizon to determine what is most conducive to accuracy and knowledge,

---

life depends on it.

then a total evidence principle might very well heed such considerations. But rationality—our rationality—has a rather different texture. It is not, on the whole, a contextual or consequentialist thing, but is rather, in many aspects at least, a local matter.

The question facing our hypothetical Eddington as he considers how much to report about the astrographic telescope's vicissitudes is not "How can I make science more efficacious as a whole?". It is: "What information is relevant to this particular argument? What are the costs to me and my readers of incorporating into my official report this particular information, here and now?". The total evidence principle says that relevant information should be provided if the cost of inclusion is low, meaning the cost of that particular information, at that particular time. It does not countenance the effect of generalizing its recommendation to any kind of information pertinent to any kind of conclusion at any time. Its recommendations are—like those of most or all of our principles of rationality—granular, particular to the immediate situation considered in isolation.[6]

To put the point succinctly, the scope of the cost-benefit analysis relevant to applying the total evidence principle is simply too narrow to excuse the iron rule's transgression of the principle on the grounds of its large-scale motivational benefits. What the iron rule recommends is therefore, in at least some instances, genuinely irrational—even though the adoption of the rule is, when contemplated in the large, considering its consequences for scientific inquiry as a whole, beneficial and therefore entirely reasonable.

5. The Coherence of Incoherence

Now, all of this might seem to engender something of a paradox. How can it be rational to engage in a process that inherently involves irrational steps?

6. For the case against epistemic consequentialism, arguing that the rules of rationality have the logically more local granularity characterized here, see Berker (2013).

Steps whose irrationality is essential to the virtues of the process as a whole, and thus to its rationality as a whole?[7]

That sense of puzzlement can be transmuted into an argument that the reasonableness of the large-scale whole—the overall reasonableness of iron rule–governed inquiry—must trickle down to the small scale in some way or other, infusing individual applications of the iron rule with rationality.

The argument I have in mind goes as follows. It is obviously rational to engage in science as a form of inquiry—it is just so very effective. The iron rule is essential to this effectiveness, so it is thereby rational to adopt, as a general policy, the iron rule. But adopting the iron rule simply means applying the rule in any of the particular cases that lie within its scope. Therefore, the application of the rule in any such case is rational. Either we must have misunderstood the principle of total evidence, or it must be overridden by some other, more powerful principle of rationality.

This piece of reasoning turns, I think, on a grand assumption about rationality that might be formulated, at least roughly, as follows:

> *Macro/micro coherence*: If pursuing a form of inquiry is rational, then all steps essentially involved in realizing that form of inquiry are rational.

The question before us, then, is whether the macro/micro coherence principle is correct, or at any rate, on the right track. I will say no, giving two counterexamples.

* * *

The first counterexample involves a Newcomb scenario. You are confronted with two boxes. You may look into either or both, taking whatever you find inside. There is a thousand dollars in the left-hand box. The right-hand box is a little trickier. An omniscient being has put a million dollars in the box just

7. The question provoking the sense of paradox might be compared to the related question of whether false beliefs can be epistemically useful (Pritchard 2017).

in case it has predicted that you will look in that box only. You know all of this. Which box or boxes should you empty?

There are two interesting options. The first, "two-boxing", is to empty both boxes. The second, "one-boxing", is to empty only the right-hand box, foregoing the thousand dollars that you know is in the left-hand box. The rationale for two-boxing is straightforward: at the time of your choice, any money is already in the box or boxes. You may as well take it all. The rationale for one-boxing is also rather tempting: the Newcomb demon is omniscient—its predictions are invariably correct—so all and only one-boxers will find the million dollars in the right-hand box.

In short, one-boxers walk away with a million dollars, two-boxers with a measly thousand. Yet most decision theorists will agree that two-boxing is the rational strategy. One way to formalize their thinking is the argument from dominance. At the time of your choice, either there is a million dollars in the right-hand box or not. If there is, you should take it, and also the money in the left-hand box. If there isn't, you should at least take the money in the left-hand box. So whatever situation obtains at the time of your choice, you'll do better if you two-box. You should, therefore, two-box.

In what follows, I will assume that two-boxing is the rational thing to do. I hope you agree.[8] If not, perhaps you can go along with this notion for the sake of the argument.

Suppose you discover that you will be faced with a Newcomb scenario in one month's time. How should you prepare? If you could use an extra million dollars, then I suggest you should do everything in your power to turn yourself, at least temporarily, into a one-boxer. That is, you should do your utmost to cultivate in yourself the disposition, when the big day arrives, to choose just the right-hand box.

One way to achieve this may be to watch footage of newly enriched one-

8. Weirich (2020), §2.5 provides a guide to the philosophical literature debating the rationality of and rationale for two-boxing.

boxers purchasing high-end Teslas, holidays in Antarctica, and so on. Another is to dwell on the arguments of the more persuasive advocates of one-boxing, rehearsing their lines of thought and of course avoiding compelling two-boxer ripostes. You might develop various strategies to put into play at the moment of choice to insulate yourself from the force of the dominance argument—dosing yourself with mind-addling substances and picturing yourself frolicking among the penguins.

I take these preparations to be entirely rational, in a straightforward way: you are taking actions which, if successful, will cause the Newcomb demon to put a million dollars into the right-hand box, from where it will make its way, delightfully, into your pocket.

Now imagine that the decisive moment has come. You make your choice—one-boxing, let's say. Was this a rational thing to do? No, it was not. Your psychological self-manipulations have successfully changed your dispositions, but they have made no difference to the logic of the situation. The dominance argument is as powerful as ever. The rational strategy is to take all the money that is available, that is, to two-box. Indeed, it is precisely because two-boxing remains rational that your preparation for the decision consisted in systematic attempts to warp your own judgment: the one-sided reading of the philosophical literature, the drugs, the cognitive technologies of distraction. The aim all along was to put yourself in a state where you would resist the power of logic and do the irrational thing.

Yet the project of inducing such a state was eminently rational. Thus, we have a large-scale project that is rational, and that has as an essential component—the culminating decision to one-box—a step that is irrational. The force of the dominance argument, which hinges entirely on the local structure of the decision situation, is undiminished by the big picture.[9] The

9. The term "local", should not be taken too literally. There is of course no spatiotemporal limit on dominance reasoning's reach. The locality is, rather, logical: the dominance argument's power depends only on the specific decision scenario to which it is applied, and not to the wider context.

macro/micro coherence principle is false.

* * *

My second counterexample to the macro/micro coherence principle is more realistic: it concerns the Cold War peace-keeping strategy called mutually assured destruction, in which potential antagonists keep stocks of weapons so large that, even if one were to launch a devastating first strike, the other would have sufficiently many weapons remaining to counter with equal devastation. Were either side to make a belligerent move, then, they would risk something close to annihilation. Thus, neither side makes such a move. Peace is secured.

I will assume that there are some circumstances in which it is rational to adopt a strategy of mutually assured destruction—presumably, a dangerously unstable world where the MAD strategy's risks, though hardly negligible, are considerably less than the even more appalling risk of the alternatives. Of course, there are those who find the strategy's ominous acronym only too apt, but I hope that you can go along with this supposition for the sake of the argument.

The power of the MAD strategy hinges on the antagonists' willingness to retaliate if attacked. Yet as the great game theorists of the Cold War, such as Thomas Schelling (1960), observed, there is a strong case to be made that retaliation is irrational—not in a narrow logical sense, but unreasonable all the same. A first strike is underway; your nation is doomed. How can it make the world a better place to wipe out your enemy as well? It is wiser, surely, to preserve a few seeds of civilization to grow something better next time around. (There are also primarily moral arguments against retaliation, of course—the targets would largely be innocent civilians—but here I focus on considerations that have the character of generic right reason.)

Again, this line of thought could be disputed; perhaps the canons of retributive justice demand that you end the world. But let me suppose otherwise. The MAD strategy is imperiled, then—like the long-term Newcomb strategy—by the prospect of a sudden rush of reasonableness at the critical moment.

Cold War strategists consequently contemplated various mechanisms to subvert such clear-eyed calculation: a mad dog in the White House, cultivation of an honor culture in the military, or total automation of the reprisal.[10]

Stepping back, we have a large-scale process, implementing the MAD strategy, that is rational, but that includes as an essential component—indeed, the central component—the cultivation of a disposition to do something irrational. Now, suppose that a terrible thing happens. The opponent launches a first strike and your own disposition to retaliate is triggered. Is the retaliation any less irrational for its being precipitated by a strategy that made perfectly good sense? Not in the least. That is at odds with the principle of macro/micro coherence.

It might be objected that the actual retaliation is not essential to the strategy in the way that the disposition to retaliate is essential. After all, the point of MAD is to bring about a state of affairs where the disposition remains unrealized; the destruction of the opponent (and perhaps the rest of the world) is merely a regrettable side effect of a scheme that did not work out as planned. That strikes me as disingenuous. A two-pronged strategy—a strategy that results in one of two moves, depending on the opponent's move—has both prongs as essential parts, however much the strategist might wish for one outcome over the other. Thus, the MAD case is a genuine example of macro/micro incoherence.

* * *

My two counterexamples suggest that the macro/micro coherence principle, for all its superficial plausibility, cannot hold across the board, and therefore

10. An alternative strategy is to convince your opponent that you've made such preparations, although you haven't. It's an uphill battle, however, to attain credibility. It may be easier, Schelling and other writers suggested, to persuade your opponent that you are somewhat capricious, and so that you *might* retaliate, than that you will retaliate for sure. Of course, there are credibility problems with this strategy, too. My remarks here apply to scenarios where those problems are solved by taking steps to cultivate a genuine disposition to retaliate, with at least some probability, and not to those in which credibility is achieved wholly through deceit.

that it cannot be a consequence of some entirely general feature of rationality. At least some principles of right reason—the total evidence principle, the dominance argument that enjoins two-boxing, and the argument against world-incinerating retaliation—have a local grain that, in determining what is rational, ignores the wider context.

Might some more restricted version of the principle, however, apply to scientific inquiry but not to the scenarios on which the counterexamples were built? Certainly, there are notable differences between scientific inquiry and the other cases. Perhaps most notable is a disparity in the relationship between the macro and the micro—between the rational long-term process and the putatively irrational small-scale steps. In the Newcomb and MAD cases, the long-term process is a slow build-up to single, crucial decision, "micro" in scale if not in consequences. In the science case, by contrast, a series of not-so-crucial "micro" decisions builds the momentum that gives the long-term process of evidence gathering its special intensity and epistemic power. Might there be a version of the macro/micro coherence principle that applies only in the latter sort of situation? Perhaps, but I leave it to my doubters to formulate and defend such a dictum. Until then, I rest with the interim conclusion that any version of the coherence principle applicable to the iron rule is flawed.

6. Conclusion

Science is a good thing, and it is reasonable and rational for society to nurture the institution of scientific inquiry and for individuals seeking to understand the universe or to make the world a better place to sign on as scientists, conforming to the iron rule and science's other norms.

At the same time, one of science's most important characteristics, its extraordinary evidence-gathering power, depends in an essential way on an irrational narrowness in the iron rule. That is quite all right. The narrowness does not compromise the reasonableness of scientists or of science as a whole. But it is undeniably a little unsettling, a little weird—or in other words,

epistemically rather extreme.

References

Berker, S. (2013). The rejection of epistemic consequentialism. *Philosophical Issues* 23:363–387.

Biagioli, M. (2012). Productive illusions: Kuhn's *Structure* as a recruitment tool. *Historical Studies in the Natural Sciences* 42:479–484.

Chandrasekhar, S. (1987). *Truth and Beauty: Aesthetics and Motivations in Science*. University of Chicago Press, Chicago.

Collins, H. M. and T. Pinch. (2012). *The Golem: What You Should Know about Science*. Second edition. Cambridge University Press, Cambridge.

Dirac, P. A. M. (1963). The evolution of the physicist's picture of nature. *Scientific American* 208:45–53.

Dyson, S. F. W., A. S. Eddington, and C. Davidson. (1920). A determination of the deflection of light by the sun's gravitational field, from observations made at the total eclipse of May 29, 1919. *Philosophical Transactions of the Royal Society of London, Series A* 220:291–333.

Earman, J. and C. Glymour. (1980). Relativity and eclipses: The British eclipse expeditions of 1919 and their predecessors. *Historical Studies in the Physical Sciences* 11:49–85.

Greene, B. (1999). *The Elegant Universe: Superstrings, Hidden Dimensions, and the Quest for the Ultimate Theory*. W. W. Norton, New York.

Hossenfelder, S. (2018). *Lost in Math: How Beauty Leads Physics Astray*. Basic Books, New York.

Johnson, G. (1999). *Strange Beauty: Murray Gell-Mann and the Revolution in Twentieth-Century Physics*. Knopf, New York.

Kuhn, T. S. (1996). *The Structure of Scientific Revolutions*. Third edition. University of Chicago Press, Chicago.

Pritchard, D. (2017). Epistemically useful false beliefs. *Philosophical Explorations* 20:4–20.

Schelling, T. C. (1960). *The Strategy of Conflict*. Harvard University Press, Cambridge, MA.

Schindler, S. (2022). Theoretical virtues: Do scientists think what philosophers think they ought to think? *Philosophy of Science* 89:542–564.

Stanley, M. (2003). An expedition to heal the wounds of war: The 1919 eclipse and Eddington as Quaker adventurer. *Isis* 94:57–89.

Strevens, M. (2020). *The Knowledge Machine: How Irrationality Created Modern Science*. Liveright, New York.

Weinberg, S. (1992). *Dreams of a Final Theory*. Pantheon, New York.

Weirich, P. (2020). Causal decision theory. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Winter 2020 edition. URL = <https://plato.stanford.edu/archives/win2020/entries/decision-causal/>.

Wilczek, F. (2015). *A Beautiful Question: Finding Nature's Deep Design*. Penguin, New York.