THE JOURNAL OF PHILOSOPHY

VOLUME LXXXVII, NO. 1, JANUARY 1990

THE DIVISION OF COGNITIVE LABOR*

BY 1804, the phlogiston theory was dead. Thirty years earlier, the same theory had been favored by almost every chemist in Europe. If the chemical revolution was resolved on the basis of reason and evidence, then it appears that there must have been some moment between 1774 and 1804 when the balance of evidence finally tipped against the phlogiston theory and in favor of Antoine-Laurent Lavoisier's "new chemistry."

Imagine that the objective degree of confirmation of the phlogiston theory just prior to noon on April 23, 1787, was 0.51, that of the new chemistry 0.49. At noon, Lavoisier performed an important experiment, and the degrees of confirmation shifted to 0.49 and 0.51, respectively. Allowing for a time lag in the dissemination of the critical information, we can envisage that there was a relatively short interval after noon on April 23, 1787, before which all rational chemists were phlogistonians, and after which all were followers of Lavoisier.

Does this scenario of initially uniform opinion, sudden jumping of ship, and new consensus signal the rational growth of scientific knowledge? If you had been a philosopher-monarch, concerned to have your scientist-subjects distribute their efforts so as to promote the eventual attainment of truth by the community, you would (rightly) have dismissed this assignment of resources (the scientists themselves) as a bad bargain. With the evidential balance between

* I am grateful to the many people who have heard or read ancestors of this paper and who have given me valuable advice, and, in particular, to John Beatty, Gerald Doppelt, Isaac Levi, David Lewis, and Elisabeth Lloyd. I owe a large debt to the ideas and the writings of Thomas Kuhn. Special thanks are due to Stephen Stich for his many detailed comments and constructive suggestions.

0022-362X/90/8701/5–22

the two theories so delicate, you would have preferred that some scientists were not quite so clear-headed in perceiving the merits of the theories, so that the time of uniform decision was postponed. A community of chemists that responded in the fashion of my original story is a badly-run community—an irrational community, if you like.

My story is intended to raise an important, if neglected,[1] problem about the growth of science. Is it possible that there should be a mismatch between the demands of individual rationality and those of collective (or community) rationality? Could it turn out that high-minded inquirers, following principles of individual rationality, should do a poor job of promoting the epistemic projects of the community that they constitute? Might those with baser motives actually do more to advance their community's epistemic endeavors? Are there conditions under which, in light of our goals as an epistemic community, we ought to want to maintain cognitive diversity? What, if anything, do we do, or can we do, about it?

I

Perhaps the predicament I describe is simply an artifact of faulty presuppositions. The following brief remarks are intended to address some obvious concerns and to indicate that the problems I have raised survive modifications of the framework I have employed in presenting them.

Any optimistic suggestion that the kind of delicate balance I have envisaged is unrealistic because scientific decisions are always clear-cut is belied by recent studies in the history of science. Moreover, even where there are large differences in measures of support, it is still possible for discrepancies between individual and collective rationality to arise. A more promising diagnosis of what goes wrong in my fable maintains that we cannot speak sensibly about numerical values representing the empirical support a theory enjoys. The more limited version of this tactic allows that degrees of support should be understood as connected subintervals of $[0,1]$. But it is easy to construct versions of my original story in which we would prefer a mi-

[1] The problem is posed by Thomas Kuhn in "Objectivity, Value Judgment and Theory Choice" [in *The Essential Tension* (Chicago: University Press, 1977), pp. 320–339]. Virtually the only subsequent treatments of the possibility of discrepancies between individual and collective rationality are the proposals of Husain Sarkar in *A Theory of Method* (Berkeley: California UP, 1983), and my own very sketchy remarks in ch. 6 of *Abusing Science* (Cambridge: MIT, 1982). I have learned much from Sarkar's treatment of the topic, even though his focus is on alternatives in methodology rather than in differences in theories, research programs, or methods.

nority of scientists to continue to espouse a theory whose objective support is measured by a subinterval of [0,0.5].

A more serious worry starts with the denial that there are objective measures of support. Does the sophisticated work in history of science not reveal to us that there are numerous cases in which equally reasonable people may disagree about the merits of rival theories, perhaps because they have different ideas about the significance of different problems or about the appropriate criteria for solving those problems? For the purposes of this essay, I do not want to take sides on this vexed question. I claim, simply, that we sometimes want to maintain cognitive diversity even in instances where it would be reasonable for all to agree that one of two theories was inferior to its rival, and we may be grateful to the stubborn minority who continue to advocate problematic ideas.

By the 1790s, only a handful of chemists continued to explore the possibility of reviving the phlogiston theory. As Thomas Kuhn[2] remarks, it was probably unreasonable for Joseph Priestley to persist as long as he did. Yet, from the point of view of the community of chemists, it was no bad thing that Priestley (and a few others) gave the phlogiston theory every last chance. Turning to our century, and to the history of Alfred Wegener's theory of continental drift, we can appreciate how things might have gone differently. In the 1920s and 1930s, Wegener's claim seemed to face insuperable difficulties, for there were apparently rigorous geophysical demonstrations that the forces required to move the continents would be impossibly large. Despite this, a few geologists, most notably Alexander du Toit, continued to advocate and articulate Wegener's ideas. I suggest that the distribution of cognitive effort was preferable to a situation in which even the small minority abandoned continental drift.[3]

Was it equally reasonable to be a drifter or an anti-drifter in the 1920s and 1930s? Inspired by appreciation of the intricate shifts in standards of appraisal that occur in the history of science, you might say, "yes." But then you face a problem of maintaining cognitive diversity of the same type as that with which I began: from the community's point of view, it would have been better if the geologists

[2] *The Structure of Scientific Revolutions* (Chicago: University Press, 1962; 2nd ed. 1970), p. 159.

[3] For an overview of the career of Wegener's theory, see Anthony Hallam, *A Revolution in the Earth Sciences* (New York: Oxford, 1973). Du Toit's persistence in supporting the theory is manifest in his book, *Our Wandering Continents* (Edinburgh: Oliver & Boyd, 1937).

had been more equally divided. On the other hand, if you accept the idea that the geophysical arguments really did expose the implausibility of Wegener's theory, then the actual distribution of cognitive effort appears better—even though, of course, some of the epistemic agents, such as du Toit, are viewed as making an irrational choice.

Consider a last suggestion for avoiding puzzles about the division of cognitive labor. We can surely distinguish attitudes that scientists adopt toward theories, hypotheses, research programs, and so forth. In particular, we can differentiate *belief* in a theory from *pursuit* of research designed to apply or extend that theory.[4] Once we have recognized the distinction, can we not accept a simple solution to my puzzle? Whereas it may be rational for each of the scientists to believe the theory that is better supported by the available evidence, it may not be rational for each of them to pursue that theory, and what the community cares about is the distribution of pursuit not the distribution of belief.

This suggested way of avoiding the discrepancy between individual and collective rationality depends on adopting two principles of individual rationality, one for belief and one for pursuit. The idea that it is rational for a person to believe the better-supported theory seems, however, to be based on supposing that that person's aim is to achieve true beliefs (or some other desirable epistemic state, the acceptance of empirically adequate theories, for example). In that case, however, it appears that the person should also pursue the better-supported theory, since pursuing a doctrine that is likely to be false is likely to breed more falsehood (or less of the desired epistemic state). Only if we situate the individual in a society of other epistemic agents—as I shall try to do in later sections—does it begin to appear rational for someone to assign herself to the working out of ideas that she (and her colleagues) view as epistemically inferior.

II

Imagine, then, that you are a philosopher-monarch, with the prerogative of directing the course of scientific research. You hope to achieve certain epistemic goals—the construction of a complete, true story of the world, the articulation of an empirically adequate theory, the elimination of error, the solving of as many problems as possible, or whatever. You have an unerring eye for detecting the objective merits of theories and complete control of the scientific

[4] See Larry Laudan, *Progress and Its Problems* (Berkeley: California UP, 1977), pp. 108–114.

workforce. What rule for the division of cognitive labor should you adopt?

Perhaps you decide to let your subjects be individually rational, allowing them to believe T whenever the objective epistemic merit of T is greater than that of its rivals (subject, perhaps, to a proviso that the merit of T be above some threshold value). But, as we saw, this is a poor strategy, liable to promote uniformity of opinion when you would prefer to keep your options open.

Even without the fiction of a philosopher-monarch, we can still consider the problem of the optimal community strategy for achieving epistemic ends. Continuing to be vague about what these ends are, I shall formulate the problem by distinguishing two types of epistemic intentions that individual scientists may have. X may have the intention that X may achieve some epistemic end (to whatever extent is possible): so, for example, X may intend that X acquire as many true beliefs as possible. This is X's *personal* epistemic intention. X may also have the intention that the community to which X belongs, the community of past, present, and future scientists, achieve an epistemic end (to whatever extent is possible): X may intend that the community of scientists uniformly adopt as many true theories as possible, in the long run. This is X's *impersonal* epistemic intention. We can recast the question "what is the rational community strategy"? as "how would scientists rationally decide to coordinate their efforts if their decisions were dominated by their impersonal epistemic intentions"? The fiction of the philosopher-monarch dramatizes the idea that this decision might require the subordination of personal epistemic intentions. If X is to engage in a community project with the goal that the community as a whole attain some epistemic end—an end that all X's fellow members also want to attain—then X may have to make decisions that do not coincide with those of an individually rational scientist. X should agree in advance that it may sometimes be necessary for some member(s) of the community to pursue (or even believe) an inferior theory, and that it may fall to X to play this role.[5]

I can now give a general description of the class of problems of optimal division of cognitive labor. Suppose that there is a set, S, of scientists, each of whom has a choice among the members of a set, R,

[5] Altruistically rational scientists are those who are prepared to pursue theories that they regard as inferior when, by doing so, they will promote achievement of the goals of their own (and their colleagues') impersonal epistemic intentions. Plainly this raises an even more bloodless ideal of scientific rationality than that criticized by historians and sociologists of science.

of rival cognitive objects. (R may be a set of rival theories, research programs, methods for approaching a problem, etc.) Each of the scientists has an impersonal epistemic intention that the community descending from S achieve some epistemic goal state: suppose that the intention of the ith scientist is that the community attain G_i (where the G_i may be different). For each of the scientists, there is an evaluation function, whose domain is R. The evaluation functions, which may be distinct, assess the epistemic merits of the members of R, and, I shall suppose, considerations of individual epistemic rationality dictate that each scientist adopt some cognitive attitude (appropriate to the category of objects belonging to R) in that member of R which ranks highest according to the scientist's evaluation function. (So, if the members of R are theories, the cognitive attitude may be belief, and the requirement of rationality may be that scientists believe that theory in R which comes out highest according to their evaluation function). The IR (individually rational) distribution of attitudes is that distribution generated among the members of S from the evaluation functions in accordance with the requirement of rationality. The CO (community optimum) distribution relative to i is that distribution of attitudes among the members of S which would maximize the probability of attaining G_i. There is a CO-IR discrepancy when there is a distribution of attitudes among the members of S which, for each i, yields a higher probability of attaining G_i than does the IR-distribution.

My general formulation allows for differences in scientists' impersonal intentions and differences in their assessments of epistemic merit, underscoring the point made in section I that problems about the division of cognitive labor arise even under the assumption that there may be changes in the standards of evaluation and changes in goals. The problems are easier to pose and easier to investigate, however, if we suppose uniformity in both respects: that is, that the G_i are all the same and that there is a single ("objective") evaluation function for each scientist. Just as I began by tacitly making this simplifying supposition, the rest of this essay will continue to adopt it. The problem only becomes more complicated if the supposition is discarded.

Given our simplification, we can talk about a community goal state, G, and an unrelativized CO-distribution. A CO-IR discrepancy will be a case in which the CO-distribution differs from the IR-distribution. I now want to descend from the abstract level of the last paragraphs, considering a particular type of problem of division of cognitive labor with the aim of identifying some conditions under

which CO-IR discrepancies can be expected to occur. Although my opening example concerned theory choice, the next two sections will explore the possibility of CO-IR discrepancies for the much more tractable case of choice between problem-solving methods. I shall return to the case of theory choice (which is algebraically more complex) in section V.

III

Once there was a very important molecule (VIM). Many people in the chemical community wanted to know the structure of VIM. Two methods for fathoming the structure were available. Method I involved using X-ray crystallography, inspecting the resultant photographs and using them to eliminate possibilities about bonding patterns. Method II involved guesswork and the building of tinker-toy models. Everybody agreed that the chances that an individual would discover the structure of VIM by using method I were greater than the chances that that individual would discover the structure by using method II. Since all members of the community were thoroughly rational, each chemist used method I. They are still working on the problem.

The community goal is to fathom the structure of VIM as quickly as possible. Suppose that each method is associated with a probability function, $p(n)$, representing the chance that the method will deliver an answer if n workers are assigned to it. Assume further that any answer delivered is recognizably either correct or incorrect. Imagine also that the relations between the probability functions represent their behavior over any time intervals we might consider—so, for any time interval, t, the probabilities that method I delivers an answer within t and that method II delivers an answer within t are in the same ratios as the functions p.[6] N workers are available for distribution between the two methods. The CO-distribution is given by having n workers use method I and $N - n$ use method II so as to maximize the probability that the structure of VIM will be discovered, that is, to maximize

$$p_1(n) + p_2(N - n) - Prob \text{ (both methods deliver)}$$

[6] These functions measure the chance that a method will deliver a correct answer, for an assignment of n workers, given that the world is as represented by the community's current knowledge about the molecule. Thus, for example, if little is known about VIM and if both methods have been pursued by similar numbers of workers for fathoming a large class of molecules, with method I proving successful much more frequently than method II, whatever the number of workers assigned, $p_1(n) > p_2(n)$. I shall not pursue here the question of exactly how to interpret the probabilities. To the best of my knowledge, it is possible to conceive of them along any of the currently popular lines.

I shall assume, for simplicity's sake, that the probability that both methods will deliver the correct answer is zero. (The effect of this—nontrivial—assumption is solely to simplify the algebra. Qualitatively similar conclusions can be obtained, at far greater length, if it is not made.)

Evidently, the solution to the problem depends on the form of the functions $p(n)$. I shall call these *return functions,* since they measure the return in probability of reaching the goal for an investment of n workers. I shall take these functions to be subject to the following constraints: they should increase monotonically with n, they should be zero when n is zero, and they should tend asymptotically to some value p when n goes to infinity (p represents the intrinsic prospects of the method, the probability of its success when we abstract from limitations of human effort).[7] Given the simplifying assumption that both methods cannot work, we know that the values of the asymptotes, p_1 and p_2, must sum to less than 1. These constraints leave a lot of room for choice of functions. To make one point explicit, it is quite possible that the forms of the functions should be different for the two methods. (Imagine that one method responds much more quickly than the other to the efforts of workers.)

I shall consider two possibilities for the functions.[8] Suppose first that $p_i(n) = p_i(1 - e^{-kn})$. Then $p_1(n) + p_2(N - n)$ is maximized when

$$n = (kN + \ln p_1 - \ln p_2)/2k$$

Notice that, even when method I has more intrinsic promise than method II ($p_1 > p_2$), there is a range of conditions—when $\ln p_1 - \ln p_2 < kN$—under which the CO-distribution is to divide the community. Intuitively, a genuine division of cognitive labor would be best for the community if there is a large available workforce (N is large), or if the methods respond quickly to the injection of effort (k is not too small), or if the difference in intrinsic promise between the methods is not too great (p_1 and p_2 are fairly close). The inequality given above represents the ways in which tradeoffs are made.

[7] I originally thought that these constraints would apply in all cases. As Stephen Stich pointed out to me, however, too many cooks may spoil the broth. Imagine, for example, that the method involves observing some sensitive organisms and that crowding in the field would disturb the organisms' normal behavior.

[8] The two classes of functions I consider represent two major possibilities: either the rate of return is fast at the beginning, and then slows as the asymptote is approached, or the rate is initially slow, speeds up once a critical mass of workers has been assembled, and slows as saturation is reached.

The functions considered in the last paragraph have the property that the rate of increase in $p(n)$ is maximal when n is small. This idea may not be at all realistic. Perhaps the chances of achieving an answer by following a given method increase quite slowly at first, then go up rapidly once a critical mass of workers has accumulated, and eventually increase very slowly as saturation is approached. We can describe this behavior by mimicking the logistic-growth equation of population biology, supposing that the p_i are given by

$$p_i(n) = p_i(3n^2 - 2n^3/kN)/k^2N^2 \qquad (n < kN)$$
$$p_i(n) = p_i \qquad (n \geq kN)$$

If the probabilities are given by these functions, then there are various cases of interest, depending on the value of k. Provided that $k < \frac{1}{2}$, it is possible to realize the intrinsic prospects of both methods, so the CO-distribution divides the workforce. If $\frac{1}{2} < k < 1$, it is not hard to show that the optimal value of n is less than kN. When k is greater than or equal to 1, n should be N (recall that method I is superior, that is $p_1 > p_2$).

Let me give a qualitative interpretation of these findings. As in the previous case, k is a critical parameter, representing the responsiveness of the methods. If the methods are so responsive that the intrinsic prospects of both can be realized with the available workforce, then it is easy to appreciate that the community epistemic interests are best served by dividing the labor. Even when k is between $\frac{1}{2}$ and 1—so that it is possible to realize the intrinsic prospects of one method but not those of both—it may be better to divide the workforce so that the prospects of neither method are realized. Provided that the difference between p_1 and p_2 is not too great, it will be better to assign a new worker to method II, if method II already has sufficient devotees to offer a large return from a new investment, rather than to method I, if method I is nearly saturated. Once k reaches 1, however, it is always better to assign all resources to the method whose intrinsic prospects are higher.

Let us now turn to the IR-distribution for these cases. On one simple understanding of individual epistemic rationality, rational agents judge methods according to the intrinsic qualities of those methods, not according to what their fellows are doing. If we understand individual epistemic rationality in this simple way, then it is easy to see that there can be discrepancies between the IR-distribution ($\langle N, 0 \rangle$) and the CO-distribution (which is sometimes $\langle n, m \rangle$ with both n, m nonzero).

Perhaps we should think of individual epistemic rationality a bit differently. Suppose it is a requirement of individual epistemic rationality that an agent maximize her chances of following a method that yields the answer. We can interpret the requirement in two ways. (1) We imagine the agent making a decision in complete ignorance of what other members of the community are doing, so that the task is to choose i so that $p_i(1)$ is as large as possible. (2) We imagine that the agent knows the current distribution $\langle r, s \rangle$, so that method I is to be chosen just in case $p_1(r+1) > p_2(s+1)$. On either interpretation, it is easy for CO-IR discrepancies to arise.

An obvious move at this point is to modify the requirements of individual epistemic rationality so that the discrepancies vanish by the magic of redefinition: simply declare that an individually rational agent is a person who chooses so as to belong to a community in which the chances of discovering the correct answer are maximized. I suggest that it is a virtue of the analysis I have been presenting that it forces into the open this altruistic ideal of rationality—an ideal that seems to me to be rather different from the concepts of rationality that figure in traditional philosophical discussions. But whether we identify community rationality and individual rationality by fiat, I am concerned with the properties of CO-distributions and the possibilities that real, imperfectly altruistic people might approximate them. The next section will explore the possibility that allowing our scientists to depart from the high-minded goals of individual rationality (and act on baser motives) might actually help the community's project.

IV

Plunging into some algebraic details, we left a community of chemists striving and failing to fathom VIM. I shall imagine that the CO-distribution for them involved a genuine division of labor (corresponding to one of those cases considered in the last section in which $p_1 > p_2$, $p_1(1) > p_2(1)$). They failed to achieve this, since all of them followed one of the principles of individual rationality which led to the distribution $\langle N, 0 \rangle$. Moreover, because the structure of VIM could only be fathomed by method II, their inability, as a community, to hedge their bets was costly.

By contrast, in a neighboring nation, the chemical community was composed of ruthless egoists. Each of the members of this community made decisions rationally, in the sense that actions were chosen to maximize the chances of achieving goals, but the goals were personal rather than epistemic. Those who elected to work on VIM did so because they believed that whoever discovered the structure of

VIM would win a much-coveted prize. Make the simplifying (but not altogether implausible) assumption that, if a method succeeds, then each person pursuing that method has an equal chance of winning the prize. How should we expect the Hobbesian community to distribute its effort?

Imagine that the community has reached a distribution $\langle n, N-n \rangle$. You are a scientist currently working on method I, and you ponder the possibility of switching to method II. The change would be good for you—given my assumption about your interests and aspirations—if it would increase the probability that you win the prize. Now the probability of your winning is the probability that someone in your group wins, divided by the number of group members. (Intuitively, by choosing a method, you buy into a lottery that has a probability of paying up, a probability dependent on the number of ticket holders; your chance of collecting anything is the probability that the lottery pays up divided by the number of tickets). Thus, at $\langle n, N-n \rangle$, it will behoove a scientist working on method I to switch to method II if

$$p_2(N-n+1)/(N-n+1) > p_1(n)/n$$

To understand how our imaginary Hobbesian agents might distribute themselves, we need to discover equilibria, points at which nobody is better off switching to the alternative method.[9] Let us say that the distribution $\langle n, N-n \rangle$ is *stable downward* if $p_1(n)/n$ is greater than or equal to $p_2(N-n+1)/(N-n+1)$, and *unstable downward* otherwise. Similarly, $\langle n, N-n \rangle$ is *stable upward* if $p_1(n+1)/(n+1)$ is less than or equal to $p_2(N-n)/(N-n)$, and *unstable upward* otherwise. $\langle n, N-n \rangle$ is *bilaterally stable* just in case it is both stable upward and stable downward.

If a community of Hobbesian scientists reaches a distribution that is bilaterally stable, then we can expect it to stay there. *Stability* is one thing, however, *attainability* another. Even though a particular distribution might be maintained, once it had been achieved, it may prove impossible for a group of self-interested scientists to reach it. For any distribution $\langle n, N-n \rangle$ that is bilaterally stable, we can define its *zone of attraction* to be the set of distributions that collapse to $\langle n, N-n \rangle$. More precisely, say that $\langle m, N-m \rangle$ collapses up to

[9] My thinking about the problem of the evolution of distributed effort in scientific communities has been heavily influenced by R. A. Fisher's classic discussion of the evolution of sex ratios [see his *The Genetical Theory of Natural Selection* (New York: Dover, 1958), pp. 158–160] and by the ideas of John Maynard Smith [particularly *Evolution and the Theory of Games* (New York: Cambridge, 1982)].

$\langle n, N-n \rangle$ just in case $m < n$, and for each x, $m < x < n$, $\langle x, N-x \rangle$ is unstable upward, and analogously for collapsing downward. $\langle n, N-n \rangle$ is *attainable* if its zone of attraction contains all distributions.

The Hobbesian community might work much better than the high-minded spirits of the last section who failed to divide the labor. More exactly, maybe there is a distribution which is both stable and attainable and which offers a higher probability of community success than the IR-distributions we considered above. (The ideal, of course, would be to show that the CO-distribution is both stable and attainable). The very factors that are frequently thought of as interfering with the rational pursuit of science—the thirst for fame and fortune, for example—might actually play a constructive role in our community epistemic projects, enabling us, as a group, to do far better than we would have done had we behaved like independent epistemically rational individuals. Or, to draw the moral a bit differently, social institutions within science might take advantage of our personal foibles to channel our efforts toward community goals rather than toward the epistemic ends that we might set for ourselves as individuals.

But is the possibility genuine? Consider cases. The simplest is that in which the return functions are given by

$$p_i(n) = p_i(1 - e^{-kn}) \text{ with } k \text{ large and } p_1 > p_2.^{10}$$

There is a bilaterally stable distribution in the neighborhood of $\langle n^*, N-n^* \rangle$, with $n^* = p_1 N/(p_1 + p_2)$. The distribution is attainable. Moreover, if p_1 is only slightly larger than p_2, the distribution yields a probability of community success that is close to that given by the CO-distribution. Moral: there are conditions under which the Hobbesians do better than their epistemically pure cousins, even conditions under which they come as close as you please to the ideal.

Life is more complicated if the return functions take the forms

$$p_i(n) = p_i(3n^2 - 2n^3/kN)/k^2N^2 \quad \text{for} \quad n < kN$$

$$p_i(n) = p_i \quad \text{for} \quad n > kN$$

$$\text{where} \quad p_2 < p_1 \quad \text{and} \quad k < p_2/(p_1 + p_2)$$

[10] It is also necessary for the claims that follow to be true that p_2 not be too small. These vague conditions can be formulated more precisely by requiring that $exp\{-kp_2N/(p_1 + p_2)\} << 1$.

Under these conditions, it is possible to achieve the intrinsic prospects of both methods and any distribution $\langle n, N - n \rangle$ with $kN < n < N - kN$ is a CO-distribution. There is a bilaterally stable distribution $\langle n^*, N - n^* \rangle$, given by $n^* = p_1 N/(p_1 + p_2)$. Provided that $p_2/(p_1 + p_2) > k$, this bilaterally stable distribution will be a CO-distribution. So far, so good. So long as the intrinsic prospects of the inferior method are not too low, and the methods respond quickly to the assignment of workers, there will be an optimal division of cognitive labor which the community can maintain—if it can but reach it.

But there is the rub. The zone of attraction of the stable distribution includes all the CO-distributions, but it is quite possible that the community should get stuck at a suboptimal distribution, particularly at the extreme $\langle N, 0 \rangle$. Intuitively, if p_2 is too small or N too large, there may be no benefit in a maverick's abandoning method I. The good news is that there are some instances of this general type in which the Hobbesian community not only does better than its high-minded cousins, but actually achieves a stable optimal division of cognitive labor. The bad news is that, when the community is too big, self-interest leads the community to the same suboptimal state as individual rationality.

There is a remedy, however. The trouble with large communities (more exactly communities for which kN is too big) is that a single deserter from method I cannot contribute enough effort to method II to make that method profitable. What is needed is for several people to jump ship together. Imagine, then, that the community is divided into fiefdoms (laboratories) and that, when the local chief (the lab director) decides to switch, the local peasantry (the graduate students) move, too. Suppose that each lab contains q members and that the director can thus bring it about that x members of the community switch where x is less than or equal to q. Of course, if $q > kN$, then a single laboratory can realize the intrinsic prospects of method II, and it is easy to see that there are conditions under which a stable CO-distribution is attainable. Moral: a certain amount of local autocracy—lab directors who can control the allegiances of a number of workers—can enable the community to be more flexible than it would be otherwise.

I have been exploring some of the consequences of a very schematic description of the ways in which personal motives (and social incentives) might operate in a scientific community. Obviously, my account would be improved by making more realistic assumptions and introducing factors that have so far been omitted (for example,

the possibility that both methods might succeed, individual differences in talent and interest, and so forth). I want to close this section, however, by taking up an obvious objection. Do my conclusions about the expected distributions achieved by the community not depend not only on the assumption that each of the members is driven by the desire to win the prize, but also on the supposition that each has enough information about probabilities to see how best to achieve that end? And is that supposition not highly unrealistic?

I reply that the supposition that scientists identify the probabilities that methods will succeed, given an assignment of a number of workers, is an idealization of the same kind as that traditionally made in confirmation theory. Just as we idealize the everyday judgments in which scientists assess the chances of competing theories, so, too, we can start from the ordinary sense that a method is overrepresented (or underrepresented) or from the awareness that the chances that a method will succeed are quite low unless it is pursued by a critical mass of people, and suppose that rough-and-ready judgments are replaced by the assignment of numerical probabilities.

Of course, appealing to human ambition is only the beginning of the story. Other psychological mechanisms might bring scientists closer to the CO-distribution than they would otherwise have been. Not only may vices from greed to fraud play a constructive role, but community ends may be furthered by more salubrious traits. Perseverance, personal investment, personal and national loyalties, and devotion to political causes may, on occasion, help to close a CO-IR discrepancy.

V

I have looked in some detail at the problem for methods. Can we achieve similar results for our original problem, the problem of theory choice?

Yes. But precise description of cases of theory choice turns out to be tricky. Crudely, the troubles stem from the existence of two sources of uncertainty: we need to take into account the probability that a theory will improve its apparent epistemic status and also the probability that, if it does so, it will be closer to the epistemic goal (e.g., truth). Nonetheless, if we are prepared to make some large idealizations, my original fable can be articulated in more detail. The text of the present section will tell a highly simplified version of the story; gestures in the direction of greater realism can be found in the footnotes.

Imagine that, at some moment in the history of some science, we have a pair of rival, incompatible theories, T_1 and T_2. Given the

available evidence, the probability that T_i is true is q_i, and all members of the scientific community concerned with the theories recognize this. Suppose further that $q_1 + q_2 = 1$, that $q_1 > q_2$, but that q_1 is approximately equal to q_2.[11]

The community goal is to arrive at universal acceptance of the true theory, to eliminate the problems that currently beset this theory, and to develop the theory both in its applications to theoretical problems and to practical matters. In pursuing this goal, the community can follow one of two generic strategies: (A) assign all the scientists to T_i; (B) assign n scientists to T_1, $N - n$ to T_2 (where $0 < n < N$). I shall consider the merits of these strategies from the perspective of a later stage in the history of the community—"the time of reckoning"—at which we assign epistemic utilities to various consequences.[12]

The possible outcomes at the time of reckoning are as follows. If everyone has been pursuing the true theory, then, I assume, that is the best of all possibilities, and, through its resolution of problems, and so forth, the community has amassed epistemic utility u_1. On the other hand, if everyone has been working on the false theory, then, I shall suppose, the effort has been completely wasted, for epistemic utility $= u_1$. To understand the consequences of dividing the labor, we need to introduce the concept of a *conclusive state*, a situation in which the present standoff between T_1 and T_2 is resolved. I shall imagine that both the available theories are currently beset by anomalies, problems that it is necessary for them to overcome if they are to win unqualified acceptance (honoring the traditional idea that theories are born refuted). We would reach a conclusive state in favor of one of the theories, say T_1, just in case, at the time of reckoning, T_1 had managed to overcome its problems and T_2 had not, despite being given an opportunity to do so. The controversy between T_1 and T_2 is resolved if *both* theories are pursued and if *one* overcomes its current anomalies, while the other does not.

Now let us make the very optimistic assumption that nature, though not forthcoming, is also not hostile: although correct

[11] There are many different idealizations here: I assume that theories can be associated with definite probabilities on the basis of the available evidence, that there is universal recognition of the right probabilities, and that one of the theories is correct. One first move toward greater realism would be to relax this last supposition, allowing that $q_1 + q_2 = r < 1$.

[12] Here it would be more realistic to allow for the possibility that the community aims to use the presently available theories in achieving a more adequate descendant theory that would be closer to the truth than either of those now available.

theories may encounter anomalies, theories that successfully overcome all their anomalies are correct. Thus, if we reach a conclusive state, then there is no need to worry about false positives—in a conclusive state we resolve the issue and we resolve it correctly. In light of this assumption, I shall assign epistemic utilities to the outcomes as follows: if division of labor by following one of the (B) strategies leads to a conclusive state, then we attain epistemic utility u_2 ($0 < u_2 < u_1$); if it does not, then the epistemic utility is 0 (we are still in the same predicament, although our labor may have given us a clearer view of the problems that each of the rivals faces).[13]

The expected utility of the more promising (A) strategy (assign all scientists to T_1) is easily computed. It is

$$q_1 u_1 - q_2 u_1$$

To work out the expected utility of the strategy (B_n) (choose the distribution $\langle n, N-n \rangle$), we need to recognize that a conclusive state in favor of T_1 will be attained just in case (a) T_1 overcomes its current anomalies, (b) enough workers are assigned to T_2 to give T_2 a chance. I shall assume that the probability that (b) is the case is 1 if $N - n$ is larger than some value, m, and 0 otherwise. The probability that (a) obtains is the probability that T_1 is true multiplied by the probability that T_1 responds to the efforts of n scientists. Letting $p_i^*(n)$ be the probability that a true theory T_i responds to the assignment of n workers by overcoming its problems, we can write the expected utility of (B_n), where $m < n < N - m$ (so that both theories are given a chance) as

$$q_1 p_1^*(n) u_2 + q_2 p_2^*(N-n) u_2$$

Division of labor is thus preferable if there is an n, meeting the "give both a chance" constraint, $m < n < N - m$, such that

$$q_1 p_1^*(n) u_2 + q_2 p_2^*(N-n) u_2 > q_1 u_1 - q_2 u_1$$

[13] My claims about the u_i can easily be adjusted to reflect differences in views about the values of particular outcomes, or even differences in specific situations which require assignment of different values. Much more tricky is the task of replacing the hypotheses about conclusive states with more realistic assumptions. In principle, one ought to allow for the possibility that ingenuity can make a false theory continue to appear plausible, and the idea of simple opposition between victory for one theory or an inconclusive dispute should give way to study of the evolution of the probabilities assigned to the rival theories. So, in effect, we want an estimate of the probability that each theory will be assigned if there is a particular division of cognitive labor, and to use this to specify the probability of making a correct decision at the time of reckoning. My preliminary explorations of the more complex algebra that results suggest that, in a significant number of instances, the much simpler approach of the text will not lead us very far astray.

I shall simply consider one of the forms of return function introduced in section III (similar results are easily generated in the other case). Suppose that $p_i^*(n) = p_i(3n^2 - 2n^3/kN)/k^2N^2$ $(n < kN)$, $p_i^*(n) = p_i$, otherwise. We can simplify the discussion without loss by supposing that $p_1 = p_2 = p$. Suppose, as with our earlier discussions, that $k < \frac{1}{2}$, so that it would be possible to assign scientists to each theory in a way that would give each a maximal run for its money. If n lies in the interval $[kN, (1 - k)N]$, then it seems reasonable to conclude that the condition that both theories have been given a chance is satisfied. So the crucial inequality for the preferability of one of the (B_n) is

$$pu_2 > (q_1 - q_2)u_1$$

Thus, unless the maximal chances of a true theory overcoming its problems are low (p is small) or the utility of immediate action is high (u_1 is large relative to u_2), the CO-distribution again involves a genuine division.

Could nonepistemic incentives operate to bring individual scientists close to the CO-distribution? Let us suppose that the important motive is each scientist's desire to be singled out by posterity as an early champion of the accepted theory. If the community is initially divided, with distribution $\langle n, N - n \rangle$, and if $kN < n < (1 - k)N$, then there is a stable attainable distribution, $\langle n^*, N - n^* \rangle$, where

$$n^* = q_1N/(q_1 + q_2)$$

Provided only that $k < q_2$ (a very weak assumption, given that $k < \frac{1}{2}$ and q_2 is close to q_1 [$= 1 - q_2$]), this will be a CO-distribution. Moral: as in the case of the earlier models, there are specifiable circumstances, albeit highly idealized, in which the IR-distribution diverges from the CO-distribution and in which extra-epistemic incentives bring the community to the CO-distribution. Social structures within the scientific community can work to the advantage of the community epistemic projects by exploiting the personal motives of individuals.

VI

I want to close with some brief indications of how the analysis I have begun here might be deepened and extended. First, it would be relatively easy to consider social structures and personal motivations (national or personal loyalties, for example) that I have left out of account. Second, in my treatment of particular cases, I tacitly assumed that the size of the available workforce was fixed. But this is surely unreasonable. Just as we can think of science as facing intra-

field optimization problems, we can also envisage issues of interfield optimization. Although personal interests may also help the community of scientists to achieve a reasonable distribution of scientists across fields, difficulties of retraining may interfere with the process.

Interfield optimization is not the end of the matter. There is a problem of division of labor of even broader scope. The epistemic goals of the community do not exhaust the set of community ends, and we can ask how, given all the aims that we have for ourselves and our fellows, we should allocate resources to the pursuit of our community epistemic goals. Given the solution to this optimization problem, we know the size of the workforce that the sciences can command. We can then ask for the optimal division of labor among scientific fields, and, finally, proceed to the question that has been addressed in a preliminary way in this essay: what is the optimal division of labor within a scientific field, and in what ways do personal epistemic and nonepistemic interests lead us toward or away from it? That question ultimately finds its place in a nested set of optimization problems.

Optimality analysis need not breed optimism. One of my main themes has been the possibility that psychological factors (and scientific institutions that exploit those factors) often thought to be detrimental to cognitive progress might turn out to play a constructive role. But it would be highly surprising if the existing social structures of science, which have evolved from the proposals of people who had quite different aims for the enterprise and who practiced it in a very different social milieu, were to be vindicated by an optimality analysis. How do we best design social institutions for the advancement of learning? The philosophers have ignored the social structure of science. The point, however, is to change it.

PHILIP KITCHER

University of California/San Diego