The Role of the Priority Rule in Science

Michael Strevens

Journal of Philosophy 100:55–79, 2003

ABSTRACT

Science's priority rule rewards those who are first to make a discovery, at the expense of all other scientists working towards the same goal, no matter how close they may be to making the same discovery. I propose an explanation of the priority rule that, better than previous explanations, accounts for the distinctive features of the rule.

My explanation treats the priority system, and more generally, any scheme of rewards for scientific endeavor, as a device for achieving an allocation of resources among different research programs that provides as much benefit as possible to society. I show that the priority system is especially well suited to finding an efficient allocation of resources in those situations, characteristic of scientific inquiry, in which any success in an endeavor subsequent to the first success brings little additional benefit to society.

1. Discovery and Priority

Science aims to provide goods that all can share, but it does so in an atmosphere that is as competitive as it is cooperative. Consider what happens when several research programs all pursue the same scientific goal: there is a competition that has the form of a winner-takes-all race, that is, a race in which there are no second prizes.

Remarkably enough, this is true in two quite distinct ways: first, with respect to the benefit that a successful research program confers on society, and second, with respect to the personal rewards—fame, prizes, authority, and so on—that accrue to the members of a successful research program. Let me say more about each.

First, benefit to society. It is a consequence of the very nature of scientific information that, while the first of several competing research programs to achieve the programs' common goal can expect to confer perhaps a great deal of benefit on society, the runners-up will confer very little or nothing. The reason is, of course, that additional discoveries of the same fact or procedure are pointless. Thus, whereas a typical economic goal, such as providing a swordfish steak or a back rub, can be realized over and over again to society's benefit—we all have use for more than one meal or massage—the goal of a scientific research program needs to be realized just once for society to benefit maximally. Competing research programs are therefore racing to be the provider of a benefit in the knowledge that the first to achieve their common goal will, in reaching it, nullify the worth of the efforts of all the other competitors. Call this the *benefits race*.

Second, personal benefit. Possibly the most distinctive feature of the social organization of science is the priority rule, the system of rewards which accords all credit, and so all the personal benefits that go along with credit, to the first research program to discover a particular fact or procedure, and none to other programs pursuing the same goal. As a consequence of the priority system, workers in competing research programs are involved in a

winner-takes-all race for personal rewards. Call this the *rewards race*.

I will argue that the strong parallel between the payoff structure of the benefits race and that of the rewards race, that is, between the scheme by which science benefits society and the scheme by which science rewards its practitioners, is not accidental: the fact that scientific rewards are distributed according to the winner-takes-all priority system, rather than some other scheme, is explained by the winner-confers-all manner in which science benefits society. The structure of the benefits race, then, explains society's implementation of the rewards race.

The form of the explanation is as follows. A scientific reward scheme such as the priority rule acts as a system of incentives, encouraging researchers to devote their time and energy to some research programs in preference to others. Different reward schemes, then, may result in different allocations of resources among competing research programs. Society has an interest in adopting a reward scheme that promotes an allocation with a relatively high expected payoff. I will show that the priority system promotes an especially efficient allocation of resources in winner-confers-all situations, that is, in situations where almost all benefit is extracted from a goal the very first time it is reached.

2. The Rule of Priority

2.1 Aspects of the Priority Rule

The reward system to which I refer as the priority rule can be divided into two parts. First, rewards to scientists are allocated solely on the basis of actual achievement, rather than, for example, on the basis of effort or talent invested. Second, no discovery of a fact or a procedure but the first counts as an actual achievement. The second part gives the rule its name, but the first part is, I think, equally worthy of note.

A qualification: it is not strictly true to say that the priority rule governs the distribution of all rewards in modern science, since most scientists receive some kind of salary, regardless of their achievement. Although it is possible to finesse the rule by holding that the basic salary of (say) an academic scientist is compensation for teaching and administration, rather than for scientific endeavor, I will for simplicity's sake put aside the question of monetary reward altogether, concentrating on rewards that take the form of reputation, a sizable office, the rapt attention of graduate students and the like—in short, the rewards that take the form of *prestige*. Prestige is, it is generally agreed, allocated more or less in accordance with the priority rule.

It was the sociologist of science Robert Merton who established the priority rule as a characteristic of the social organization of science worthy of serious study.[1] Merton's discussion, still an essential source for any research on priority, seeks to establish a number of claims about the priority rule of which I single out three as especially important.

First, Merton argues that the concern for priority and for actual achievement is not so much a consequence of the psychological makeup of scientists as the reflection of a powerful norm governing the scientific endeavor. Scientists, then, are not merely psychologically disposed to allocate prestige in accordance with the priority rule; they feel obliged to do so—they regard a distribution of prestige in accordance with the rule as right and proper, and a distribution that violates the rule as wrong or unjust. So, for example, Merton holds that disputes over priority are to be explained principally as a consequence of the moral concern of the scientific community that the rules be enforced, not as a consequence of a clash of individual egos.

Second, Merton documents a concern with the priority rule going back to the beginnings of modern science, that is, to the early seventeenth cen-

1. "Priorities in Scientific Discovery," *American Sociological Review* XXII (1957): 635–659, reprinted in *The Sociology of Science* (Chicago: Chicago University Press, 1973), 286–324. Page references are to the latter version.

tury. Galileo and Newton are cited as early and eager parties to priority disputes that arose from a concern with the rule; according to Merton, the rule has held sway ever since. The priority system, then, seems always and everywhere to have determined the allocation of credit in Western science.

The third Mertonian observation that I wish to consider here concerns the extreme literalness with which the priority rule is enforced: if the same fact is discovered twice, Merton notes, the first discovery garners all the rewards no matter how slender the margin by which it edges out the second. To this effect, Merton quotes the permanent secretary of the French Academy of Sciences and distinguished physicist François Arago, writing in the first half of the nineteenth century:

> Questions as to priority may depend on weeks, on days, on hours, on minutes.

This attitude, Merton writes, "only [puts into] words what many others have expressed in behavior".[2] He cites a dispute between Galileo and Simon Mayr over the observation of Jupiter's moons in which the relevant time interval was a mere ten days.[3]

Call this phenomenon—the fact that no finish is so close that the priority rule ceases to apply—the *Arago effect*. Merton regards the Arago effect as a pathology of the social organization of science, a "dysfunctional extreme far beyond the limits of utility".[4] His view is shared by others writing on the priority rule, either explicitly, or implicitly in the sense that their explanations of the rule do indeed accord no advantage to a concern with days, hours, minutes. A feature of my own explanation of the priority rule is that it does find a constructive role for the Arago effect to play: worrying about a matter

2. "Priorities in Scientific Discovery", p. 322.
3. "Priorities in Scientific Discovery", p. 287.
4. "Priorities in Scientific Discovery", p. 322.

of ten days between discoveries can exert a positive influence on the way in which science is conducted.

2.2 Alternatives to the Priority Rule

There are a number of salient alternatives to the priority system; three will provide sufficient contrast for my purposes here.

1. Scientists might all, on an egalitarian reward system, receive equal compensation, that is, for the purposes of this paper, equal prestige.

2. Scientists might be compensated in proportion to the talent and effort they invest in their research. Harder working, cleverer scientists would receive more prestige.

3. Scientists might be compensated in proportion to their actual achievements, in accordance with the priority rule, but with the amendment that being the second to make a discovery, or having progressed considerably towards making a discovery, count as actual achievements. Runners-up in a scientific race would, in other words, receive prestige for finishing well.

The question with which I am concerned, then, may be put thus: why does the scientific community disburse prestige in accordance with the priority rule rather than one of the above reward schemes, or some alternative scheme?

The aspect of the question that is perhaps most puzzling concerns the contrast between system (2), which rewards participants according to talent and effort invested, and the priority rule and system (3), which reward participants according to actual achievement. I say this because the factor that makes the difference between talent and effort, on the one hand, and actual achievement, on the other, would appear to be simple luck. Two equally brilliant scientists invest equal time in pursuing some goal; one happens to

choose the right method and so succeeds, the other does not. What reason can there be to concentrate all rewards on the scientist who makes the lucky choice? Any truly satisfying explanation of the priority rule must answer this question.

With an eye to this problem and others, I propose the following three-way classification of reward systems. First, there are egalitarian schemes, which distribute rewards more or less equally, such as scheme (1). Second, there are what I will call resource-based schemes, which distribute rewards in proportion to the resources invested by the rewardee. Scheme (2) is a resource-based scheme. Third and finally, there are achievement-based schemes, which distribute rewards in proportion to the achievements of the rewardees. Both scheme (3) and the priority rule are achievement based. This classification is not intended to be exhaustive; in section 3, I will describe a system that seems to fit somewhere between resource- and achievement-based schemes.

The main questions I seek to answer in this study can now be posed as follows:

1. What explains the fact that science's reward system is achievement based rather than resource based?

2. Given that science has an achievement-based system, why do second and subsequent attainings of a goal not count, for the purpose of determining rewards, as achievements?

2.3 Explanations of the Priority Rule

Merton and others after him have explained the primacy of the priority rule in a number of ways. Let me sketch four such explanations.

1. The priority system gives scientists an incentive to undertake research, both by rewarding them directly for their research and by creating the

perception, through the rewards, that discovery has its own intrinsic value.[5]

2. It is much easier to measure achievement than to measure the talent and effort invested in a research program. Furthermore, it is difficult to assess the progress made by the runners-up once a scientific race has been won, since it is easy for runners-up to steal the now-published results of the winners. Thus, the priority system is far more practical than reward systems such as schemes (2) and (3) above.[6]

3. The priority rule gives scientists an incentive to publish their research as soon as possible, making the benefits promptly available to society.[7]

4. It may simply seem fairest to reward scientists in exact proportion to the benefits they confer on society, perhaps especially so to the members of a market-driven society.

All four explanations have their merits. They also have drawbacks, but this is not the place to investigate the drawbacks. None of the four can, by itself, explain all aspects of the priority rule considered above. This advantage I claim for my explanation alone, to which I now turn.

3. Resource Allocation: The Additive Case

3.1 The Problem of Resource Allocation

Every scientific research program might usefully be better funded, better staffed, better equipped. When new resources, pecuniary and human, come

5. Merton, "Priorities in Scientific Discovery."

6. Partha Dasgupta and Paul A. David, "Toward a New Economics of Science," *Research Policy*, XXIII (1994): 487–521.

7. Dasgupta and David, "Toward a New Economics of Science."

available for science, there will always be a real question as to how best to distribute them among research programs, both those that exist and those that ought to exist. This is the problem of resource allocation.

There is a wide and a narrow sense of the problem. In the wide sense, all research programs are competing for money and personnel. But especially fraught—and this creates the problem in its narrow sense—is the competition between research programs aiming to realize the same goal. This paper considers both problems, first the problem of distributing resources among research programs with different goals, in this section, and then the problem of distributing resources among programs competing to reach the same goal, in section 4.

It is the contrast between the solutions to the two kinds of problems that will provide the basis for my explanation of the priority system. Any reward scheme functions as a system of incentives, directing researchers towards some programs rather than others. A reward scheme, then, has a considerable effect on the allocation of labor among research programs. I will show that the priority system is apt to achieve an especially efficient allocation of labor in a winner-confers-all benefits race, and so an especially efficient allocation of labor among research programs competing to achieve the same goal. (Readers seeking a better sense, ahead of time, of how the argument goes might turn to section 4.2.)

3.2 A Simple Model of the Resource Allocation Problem

The question of resource allocation in science was raised quite some time ago by Charles Sanders Peirce;[8] it has recently been brought once again to

8. "Note on the Theory of the Economy of Research", in Arthur W. Burks, ed., *Collected Papers of Charles Sanders Peirce* volume 7 (Cambridge MA: Harvard University Press, 1958), pp. 76–83.

philosophers' attention by the multifaceted investigations of Philip Kitcher.[9] My simple mathematical model for studying resource allocation is borrowed, along with the idea that there might be a connection between resource allocation and reward schemes, from Kitcher.

The starting points for the model are the following assumptions, which are of course quite idealized:

1. Every research program has a single goal. There are only two possible outcomes of the program's endeavors: total success, if it realizes the goal, or total failure, if it does not.

2. Different research programs have different intrinsic potentials.

3. A program's chance of success—that is, the probability that it will achieve its goal—depends on two things, its intrinsic potential and the resources invested in the program.

Given these assumptions, a research program's probability of success can be written as a function of the resources invested in it. I call this function the program's *success function*, and I write it $s(n)$, where n quantifies the resources invested.

One research program has more intrinsic potential than other if, given any fixed level of investment, the one has a higher chance of success than the other. For simplicity's sake, I assume that there is a complete ordering of programs in terms of their intrinsic potential, or equivalently, that the success functions for two programs do not intersect unless they entirely coincide. This is also true of Kitcher's model.

The problem of resource allocation is the problem of distributing limited resources among different programs so as to maximize the return to society. As Peirce and Kitcher have noted, it is not in general true that the optimal

9. "The Division of Cognitive Labor," this JOURNAL, LXXXVII, 1, (January 1990): pp. 5–21 and *The Advancement of Science* (Oxford: Oxford University Press, 1993).

allocation devotes all resources to the program with the greatest intrinsic potential; the problem, then, is not trivial.

In order to pose the problem more precisely, it is necessary to write down a function expressing the value to society of any given distribution of resources. To this end, suppose that for any research program, there is some fixed utility to society of that program's success achieved in isolation. Writing the utility v, the expected return from investing n resources in a program is the product of the utility of success and the probability of success, that is, $vs(n)$.

Now suppose that there are two research programs, with utilities v_1 and v_2, and success functions $s_1(\cdot)$ and $s_2(\cdot)$. What is the expected return to society from investing n_1 resources in the first and n_2 resources in the second? In the simplest case, it is the sum of the individual expected returns, namely,

$$v_1 s_1(n_1) + v_2 s_2(n_2)$$

I call this the *additive case*. All other cases are *nonadditive*.

Let me give an example of an additive case and an example of a nonadditive case. Expected returns from two programs will be additive if the goals of those programs are, in an intuitive sense, independent. If one program seeks a cure for cancer and another the Higgs boson, for example, then benefits are independent and returns are additive.

Returns are nonadditive in many circumstances, but the most important for my purposes is the case in which (a) both programs have the same goal, and (b) second and subsequent realizations of that goal are of no value to society, creating a winner-confers-all benefits race. In the winner-confers-all case, the expected return is the utility v of the goal in question multiplied by the probability that at least one program succeeds, that is,

$$v(s_1(n_1) + s_2(n_2) - s_{12}(n_1, n_2))$$

where the third term $s_{12}(n_1, n_2)$ is the probability that both programs achieve their goals, given the allocation of n_1 resources to the first and n_2 resources

to the second. It is the existence of this third term that makes the expected returns nonadditive.

Because I am interested in winner-confers-all races, it is the nonadditive case that ultimately concerns me. However, as I said above, what concerns me more specifically is the contrast between the optimal allocations of resources in the additive and the nonadditive cases. For this reason, I investigate the optimal allocation for the additive case in the remainder of this section.

3.3 Optimality in the Additive Case

You have a limited number $\mathcal{N}$ of resources—say, worker-hours—to distribute between two research programs. Your goal is to maximize the total expected return. The programs' benefits are additive. Assuming that additional worker-hours always improve a program's chances of success, you will distribute all $\mathcal{N}$ worker-hours available, n to one program and $\mathcal{N} - n$ to the other. Your return, then, is

$$v_1 s_1(n) + v_2 s_2(\mathcal{N} - n).$$

The problem is to find the value of n that maximizes this return.

For simplicity's sake, assume that v_1 and v_2 are equal. (Alternatively, assume that the utilities are built into the success functions, so that $s(n)$ represents not the probability of success but the expected return, given an investment of n worker-hours.) Then what must be maximized is the expected number of successes:

$$s_1(n) + s_2(\mathcal{N} - n).$$

This is the kind of constrained maximization problem that is familiar from economics. In solving the problem, I make one very important assumption, that the success functions yield *decreasing marginal returns*, that is,

that each additional worker-hour invested in a program increases the probability of success a little bit less than the last (though there is always some increase, as assumed above).

The notion of a marginal return on investment will be an important one in this paper, so let me present a definition and some notation to smooth the discussion. The marginal return from investing one worker-hour in a research program, given that n worker-hours have already been invested, is the increase in the probability of success brought by the investment of that additional hour, that is,

$$s(n+1) - s(n).$$

I write this as $m(n)$. To say that marginal returns are decreasing is to say that the function $m(n)$ is strictly decreasing, that is, that the graph of $m(n)$ slopes downwards from left to right.

The assumption of decreasing marginal returns is a familiar one; I will not try to justify it here (though I note that it may have its exceptions, as suggested by Kitcher).[10] From an economic point of view, the assumption has a pessimistic feel: a constant rate of progress requires ever-increasing rate of investment. But from a mathematical point of view, decreasing marginal returns are a wonderful thing, because the assumption of decreasing marginal returns makes possible extremely general claims about constrained maximization problems such as the resource allocation problem.

In the additive case, the assumption of decreasing marginal returns implies that the function to be maximized will have a single maximum point corresponding to the value of n at which the marginal returns are equal, that is, at which

$$m_1(n) = m_2(\mathcal{N} - n).^{11}$$

10. "The Division of Cognitive Labor," p. 13.

11. Because the marginal return function is an approximation to the derivative of the success function.

It is not guaranteed that this value of n is realistic, that is, nonnegative, but provided that N is large enough—and I assume throughout that it is—the guarantee can be made. The result generalizes in a number of ways, some of which are described at the end of section 3.4.

If the additive version of the resource allocation problem is thought of as a static problem in central planning, then, its solution is straightforward. Simply allocate the N worker-hours available to you, the central planner, so that the marginal returns from each of the programs, on the investment of a further worker-hour, would be about equal.

The same solution can be used for dynamic central planning, that is, central planning where a constant stream of worker-hours is coming available, and must be allocated as they arrive so as to provide as high a possible return on investment over time.

To see this, note that returns over time will be maximized if, for any N, over the interval where exactly N worker-hours, and no more, have become available, they are distributed in such a way as to maximize returns. Provided that the central planner always distributes worker-hours so as to keep the marginal return functions as equal as possible, the static result guarantees that this condition will be satisfied for any N, and so the dynamic problem is solved.

This treatment of the dynamic problem makes two rather significant assumptions, first, that a program's success function, and in particular, its intrinsic potential, does not change over time, and second, that the central planner knows at all times the true form of the success function. These assumptions are obviously oversimplifications; the issue is further discussed at the end of section 3.4.

3.4 A Reward Scheme for the Additive Case

In the last section, I imagined that resource allocation is to be administered by a single, all-powerful central planner who has only the interests of sci-

ence and society at heart. But in fact, with respect to worker-hours at least, allocation in science is driven to a great extent by certain decisions of individual scientists, namely, their decisions as to what projects to pursue. These decisions are, I presume, strongly influenced by scientists' expectations of rewards, in particular, the complex of rewards that I am calling prestige; it is at this point that reward schemes at last make their entrance.

Let me simplify the issues by making the following two assumptions: first, that the distribution of worker-hours among research programs is entirely determined by the choices of the workers themselves, and second, that the workers' choices are entirely determined by the prevailing reward system, so that they choose to allocate their time to the project that will bring them the greatest expected return. I am imagining, then, that every researcher periodically—at the end of a subproject, say—reconsiders their commitment to their current program, staying on only if the reward for devoting their next, say, year to that program is at least as great as the reward for investing the year in some other program. If they do not stay, they move, of course, to the program that will reward them best for their year's service.

It is important to note that the programs themselves do not distribute the relevant rewards. The programs may pay salaries, but they have no power to bestow prestige. That is a power of the scientific community as a whole, to be exercised in accordance with the prevailing reward system, an aspect of what Merton calls the "normative structure of science".

From the perspective of the problem of resource allocation, an ideal reward system is a scheme of rewards that encourages scientists to allocate their labor in accordance with the precepts laid down in the last section, that is, so as to keep the marginal returns from every program as close as possible.

This can be done, obviously enough, by encouraging scientists always to invest their time in the program that, at the moment of their decision, offers the greatest marginal return. Given decreasing marginal returns, this

will eliminate, as quickly as possible, the differences between the marginal returns from different programs. An optimal reward scheme, then, is one that, at any given time, makes the program with the highest marginal return the most attractive.

The simplest such scheme awards prestige in proportion to marginal return. That is, it bestows prestige in proportion to the amount by which a scientist increases their program's probability of success, or more generally, the amount by which they increase their program's expected utility to society.[12]

Call this scheme the *Marge* reward scheme. *Marge* is not an achievement-based scheme, because it rewards contributions to the probability of success rather than successes themselves. But it is not a purely resource-based system either, or at any rate, scientists' rewards are not determined solely by the talent and effort that they invest.[13]

---

12. Let me situate this reward scheme in a more general framework. Define as a program's *reward function* the function $r(n)$ that yields the reward given to a person who invests a worker-hour in the program at the point at which n worker-hours have already been invested. For example, if you join a program in which i worker-hours have been invested, you receive the equivalent of $r(i)$ utility points. Suppose that there are a number of different programs, each with its own reward function, and suppose that each worker always invests their next hour in such a way as to maximize their reward. Then, provided that the reward functions are *strictly decreasing*—that is, provided that the more worker-hours have been invested in a program, the lower the reward to a worker who invests one more—workers will distribute their investment among the programs so that, at any particular time, the disparity between the rewards that would be obtained by investing one more hour in any of the programs is as small as possible, that is, so that the reward functions are, at any time, approximately equal. To ensure that marginal returns are kept approximately equal, then, set the reward function for a program proportional to its marginal return function. More generally, to ensure that any other set of functions, such as the adjusted marginal return functions to be described in section 4.1, are kept approximately equal, use those functions as the reward functions.

13. My model does not represent talent explicitly, but it is easy to build talent into the representation by counting an hour of work from a talented scientist as worth more than

In addition to talent and effort, what is rewarded is timing. Under a regime of decreasing marginal returns, an hour invested earlier in a program's life brings a greater marginal return than an hour invested later. The reward for the earlier hour will be commensurately greater than the reward for the later hour. Thus, two scientists of equal talent and industry, working within the same program, will not be rewarded equally. The scientist whose career unfolds earlier in the program's history will make a greater contribution to the program's probability of success, and so will be rewarded with greater prestige than their psychological double.

For expository purposes, I nevertheless classify *Marge* as a resource-based reward scheme, on the grounds that rewards are based entirely on a scientist's input, regardless of the output. Think of the value of the resources provided by a scientist as determined not only by the intrinsic properties of those resources—by talent and effort—but also by timing.

In summary, then, there is a reward system, the *Marge* scheme, that will ensure an optimal allocation of labor among any set of research programs, provided that the programs have goals with additive values and yield decreasing marginal returns on investment. This is a result of considerably greater generality and precision than has previously been stated in the literature on resource allocation in science.[14]

But the result has its limitations, due to the assumptions and idealizations I have made along the way. Some of these can be relaxed without compromising the result. I have already noted that the result is true for cases in which there are more than two competing research programs, and that the

---

one "standard" hour.

14. Kitcher, for example, examines only one particular kind of success function, and suggests a reward scheme—an equal division of some fixed prize among the participants in a successful research program—that will normally result in a somewhat suboptimal allocation of labor. Kitcher's main contribution is not a particular mathematical result, but rather the insight that different scientific reward systems will result in different solutions to the resource allocation problem, a sine qua non of my own work.

effect of talent is easily incorporated into the model (see note 13). Two other respects in which the assumptions can be loosened without lessening the optimality of the *Marge* scheme are as follows.

First, it is possible to assign to each program a number of goals, rather than a single goal, each with its own probability of realization. Provided that these goals are additive, and that scientists are rewarded in proportion to the sum of their contributions to the realization probabilities of each of their program's goals, *Marge* will find the optimal allocation of labor.

Second, it is possible to incorporate the effect of positive feedback among different research programs, such that an investment in one program increases the probability not only of that program's success but also of other programs' successes. Provided that a scientist's reward recognizes their contribution not only to their own program, but to all other programs as well, *Marge* will steer scientists so as to achieve the best possible allocation of labor.

There are other assumptions that cannot be relaxed without compromising *Marge*'s optimality. If, for example, the success function of a program can change in unforeseeable ways, optimality may be out of reach. But this is a state of affairs that equally hinders the wisest central planner. Because investment of time is irrevocable—hours spent in research cannot be retrieved and reassigned—it is always possible that a change in a success function will create a situation in which previous resources turn out to have been invested so badly that it is impossible, with the resources in hand, to attain the optimal allocation. The problem, then, is not a consequence of any peculiar fragility of the reward scheme that I have suggested, but is rather a universal obstacle to achieving the best possible outcome when information is limited.

Indeed, the *Marge* strategy will do as well as a central planner could under the same circumstances: it will move to accommodate unanticipated changes in the success functions as fast and efficiently as possible. Although *Marge* may not achieve the optimal allocation of resources in such a case,

then, the allocation is as good as can be expected given the information available. The same comment applies to any case in which knowledge of the success function is limited.

These are perhaps not very exact claims, but precision is, in any case, about to be abandoned. The investigation of winner-confers-all races in the next section will require a move from absolute statements of the form *This reward scheme is optimal* to comparative, qualitative statements of the form *Of these two reward schemes, this one is better*. It is the comparative, qualitative statements that will form the basis for the explanation of the priority rule.

4. Resource Allocation: The Winner-Confers-All Case

4.1 The Winner-Confers-All Benefits Race

In a classic scientific winner-confers-all benefits race, two research programs compete with one another to make the same discovery. When one program succeeds, society receives the full benefit of the discovery; whether or not the other program later also succeeds makes no difference to society's welfare.

As stated in section 3.2, society extracts the greatest return from a winner-confers-all race when the allocation of resources among programs maximizes the probability that at least one program succeeds. If it is assumed, for expository simplicity, that the successes of two competing programs are independent (nothing important will turn on this assumption), then what must be maximized is

$$s_1(n_1) + s_2(n_2) - s_1(n_1)s_2(n_2).$$

The optimal distribution of N worker-hours, then, is that in which the number n of worker-hours allocated to the first research program maximizes

$$s_1(n) + s_2(\mathcal{N} - n) - s_1(n)s_2(\mathcal{N} - n).$$

Define the marginal return function $m(\cdot)$ for a program as before, so that $m(n)$ is the increase in the probability of success due to an additional worker-hour being invested on top of n hours already invested. Thus, $m(n) = s(n+1) - s(n)$, as in the additive case. Then the optimal value of n will be the value for which

$$m_1(n)(1 - s_2(\mathcal{N} - n)) = m_2(\mathcal{N} - n)(1 - s_1(n)).$$

I call the functions on either side of the equals sign the *adjusted marginal return functions*.

What I want to emphasize about this solution to the winner-confers-all allocation problem is not its exact form, which after all depends on the perhaps dubious assumption that the success of one program is independent of that of the other, but rather its qualitative properties. Specifically, what is important is that, compared with the optimal distribution for the additive case, the optimal distribution for the winner-confers-all case allocates more workers to the program with the higher intrinsic potential.[15] This observation is generally true, both for cases where more than two research programs compete to achieve the same goal, and for cases where the successes of different programs are not independent.[16] (To avert any confusion: the optimal distributions for both additive and winner-confers-all scenarios allocate more resources to the program with the higher intrinsic potential. The distribution for the winner-confers-all scenario, however, favors the higher-potential program more heavily than does the distribution for the additive

15. To see this, note first that the greater the relative value of a program's marginal return function, the more resources are allocated to that program, and second, that compared to the marginal return functions, the adjusted marginal return functions will be relatively lower for programs with relatively lower intrinsic potentials, due to the adjusting factor's negative dependence on the success probability of the competing program. (By the adjusting factor, I mean the $(1 - s_2(\mathcal{N} - n))$ factor.)

16. There are exceptions to this latter claim in extreme cases, but the extreme cases are not realistic.

scenario.)

It follows that the optimal distribution in a winner-confers-all race is different from the optimal distribution in an additive scenario. Because the optimal distributions differ, the *Marge* reward scheme will not be optimal in the winner-confers-all case. A new reward scheme is needed.

4.2 Three Reward Schemes

The optimal distribution of resources in a winner-confers-all benefits race allocates relatively more resources than the *Marge* scheme to a program with higher potential. I now show how the *Marge* scheme can be transformed so as to create the necessary additional bias in favor of higher-potential programs. The effect of the transformation is to turn *Marge* into science's priority system.

The transformation has two steps. First, *Marge*'s resource-based scheme is transformed into an achievement-based scheme on which scientists are rewarded only if their program achieves its goal. Priority plays no role in this scheme: if two rival programs both achieve their common goal, scientists in both programs are rewarded. I call this scheme (or rather, the specific version of the scheme that I describe in section 4.3) the *Goal* reward scheme.

The second step of the transformation turns *Goal* into a scheme in which there is a priority race, so that, if two rival programs both achieve their goal, only the first to do so is rewarded. I call the resulting scheme the *Priority* reward scheme.

The two new schemes will retain *Marge*'s principle for dividing rewards among a program's workers: scientists are rewarded in proportion to their contribution to their program's probability of success. What changes are the rules for deciding when programs are to be rewarded at all. The reward schemes become progressively less generous: *Marge* rewards all programs, *Goal* rewards only programs that achieve their goals, and *Priority* rewards only programs that are the first to achieve their goals.

In section 4.3, I will show that *Goal* favors higher-potential programs relatively more than does *Marge*, that is, that it attracts relatively more workers to higher-potential programs. In section 4.4, I will show that *Priority* favors higher-potential programs relatively more than does *Goal*. Moving in the direction of a priority system, then, affects the allocation of resources so as to better suit a winner-confers-all benefits race. Whether *Goal* or *Priority* achieves a more efficient allocation in a winner-confers-all race is not, however, a question that I will be able to answer here, for reasons to be explained in section 5.1.

4.3 From a Resource to an Achievement Basis

The *Marge* scheme is transformed into the *Goal* scheme by rewarding only those programs that achieve their goal. I will, in this section, make one further assumption about the workings of *Goal*, as a consequence of which it will turn out that a scientist's expected reward from any investment is the same on the *Goal* scheme as on the *Marge* scheme (down to a constant of proportionality). Yet *Goal* achieves a different allocation of resources from *Marge*, I will go on to argue, because of human *risk aversion*.

The further assumption I make about *Goal* is that there is a fixed amount of prestige bestowed on each program that is rewarded, to be divided among the program's workers according to their contributions. A scientist's share of this prize, then, is equal to their probabilistic contribution to their program's success as a proportion of the program's total probability of success. This means that, although workers in programs with relatively lower intrinsic potentials have a relatively lower chance of receiving any reward at all, when they are rewarded, they will receive a relatively larger share of the prize, since lower-potential programs have lower success probabilities.

These two factors exactly cancel out, so that on *Goal*, the expected reward for a scientist's investing in a particular program is proportional to the probabilistic contribution that the investment makes, in absolute terms, to

the program's success (with the constant of proportionality the same for all programs). In other words, the expected value of investing in a program is proportional to the reward that the scientist would receive on the *Marge* scheme for the same investment. As a consequence, if scientists were to make investment decisions solely on the basis of the expected reward from their investments, *Goal* would result in exactly the same preference ordering among programs in all situations as the resource-based *Marge* scheme. Either scheme, then, would result in the same distribution of worker-hours among research programs.

To undermine this unwelcome conclusion, I appeal to risk aversion. I use the notion of risk aversion in its technical sense. An agent is risk averse with respect to some kind of resource, such as money or social influence, if that resource has *decreasing marginal utility* for the agent, meaning that, when the agent has a certain quantity of the resource, a small amount more is not worth as much to them as it would be if they had less of the resource. The qualitative relation between the quantity of the resource and the utility is shown in figure 1.

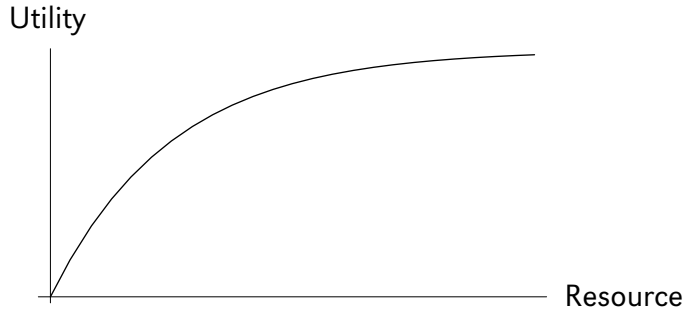

Figure 1: The relationship between the utility and the quantity of a resource, for an agent who is risk averse with respect to the resource

If a person is risk averse with respect to, say, money, they will think and behave in certain characteristic ways. For example, they will consider a

loss of a given value to be more undesirable than a gain of the same value is desirable. They will reject some fair gambles. And, most important for what follows, when two equally priced lottery tickets have the same expected monetary value, but one has a higher probability of winning (which means it must have a lower payoff), a risk averse decision maker will prefer the higher probability ticket. For example, given a choice between playing a game in which, if a tossed coin lands heads, you win $1000 (there is no gain or loss if the coin lands tails), and a game in which, if the single red ball is drawn from an urn containing 100 balls, you win $50,000, if you are risk averse, you will choose the coin toss game. To put it qualitatively, you prefer a large probability of winning a small prize to a small probability of winning a large prize.

It is this preference that allows an achievement-based reward scheme to increase the relative appeal of higher-potential research programs. Consider a scientist choosing between two rival research programs that have different potentials but that are currently offering equal expected rewards. On the *Marge* scheme, the scientist is sure to be rewarded whichever program is chosen, so they will be indifferent between the two. On *Goal*, the scientist is more likely to be rewarded by the higher-potential program than by the lower-potential program. A risk averse scientist will therefore choose the higher-potential program. For this reason, more scientists will join higher-potential programs on *Goal* than on *Marge*, as desired in a winner-confers-all benefits race.

4.4 Implementing a Priority Race

What effect will moving to a priority race have on an achievement-based scheme such as *Goal*? To answer this question, consider the relation between a scientist's expected reward on the *Goal* scheme and the same scientist's expected reward on *Priority*, a reward scheme identical to *Goal* except that it rewards only the first program to achieve any given goal.

Suppose, in particular, that a scientist is choosing between two research programs that would seem equally attractive on the *Goal* reward scheme. The first of these programs has the higher intrinsic potential, let us say, and so a higher probability of reaching its goal, but the second has few enough workers that the additional share of the reward a scientist in that program would receive, were the program to realize its goal, exactly makes up for the lower probability of being rewarded. (Assume that the effects of risk aversion are also taken into account.)

How would the scientist's preferences change if the *Goal* scheme were to be replaced, at this moment, by the *Priority* scheme? The change will not affect the size of the reward the scientist is liable to receive. Its impact is confined to the probability of the scientist's being rewarded.

In a two-program race, *Goal* rewards a scientist in three circumstances:

1. The scientist's program achieves its goal but the other program does not.

2. Both programs achieve their common goal; the scientist's program achieves the goal first.

3. Both programs achieve their common goal; the scientist's program achieves the goal second.

The *Priority* scheme rewards the scientist in situations (1) and (2) above, but not in situation (3). Because the situations are mutually exclusive, the probability of a reward on *Priority* is equal to the probability of a reward on *Goal* less the probability that situation (3) occurs.

Upon moving to *Priority*, then, the probability of a reward will decrease for both programs. The amount of the decrease is the probability of situation (3), which is $s_{12}w_2$ for the first program and $s_{12}w_1$ for the second program, where s_{12} is the probability that both programs achieve their goal (given the allocation of resources at the time of the decision), and w_1 and w_2

are the probabilities that the first and the second programs respectively will win the priority race, in the event that both achieve their goals.

It seems reasonable to suppose that, in almost all cases, the program with the higher probability of achieving its goal—in the example, the first program—is also more likely to win a priority race. Thus, w_1 is greater than w_2. It follows that, upon a move from *Goal* to *Priority*, the probability of the first program's being rewarded decreases by less than does the probability of the second program's being rewarded. Consequently, the expected reward from an investment in the first program will decrease by less than the expected reward from an investment in the second.[17] The scientist who is undecided under *Goal* will therefore choose the first program—the program with the higher potential—under *Priority*. (Risk aversion exacerbates this effect.) More generally, a move from *Goal* to *Priority* has the effect of increasing the bias in favor of higher-potential programs.

To summarize, in section 4 I have shown that

1. To find the optimal allocation of resources in a winner-confers-all benefits race, a greater bias toward high-potential programs is needed than is created by *Marge* (section 4.1), and

2. *Goal* makes higher-potential programs relatively more attractive than *Marge* (section 4.3), and *Priority* makes them more attractive still (section 4.4).

5. The Priority System Explained

5.1 From Additive to Winner-Confers-All Cases

I make the following conjecture: a resource-based reward scheme that distributes labor optimally in resource allocation problems that are additive,

17. The more so because the size of the second program's reward is larger.

or nearly additive, can be adapted to work well in winner-confers-all benefits races by retaining the scheme's apparatus for calculating rewards while moving to an achievement basis for determining the recipients of the rewards. In a winner-confers-all race, then, figure rewards as before, but give the rewards only to programs with concrete achievements.

This, I suggest, is the explanation, or at least a great part of the explanation, of the priority system in science. Our normal reward schemes, devised to handle problems of resource allocation that are close to additive, are modified for scientific resource allocation by adopting something similar to the previous section's *Priority* scheme.

The proposed explanation provides very satisfying answers to two questions about the priority system posed at the beginning of this paper concerning, first, the apparently excessive role of luck in the priority system's selection of reward recipients, and second, the status of the Arago effect.

The question about luck is raised by the observation that two equally talented, equally industrious scientists may receive different rewards under the priority system just because one of them, but not the other, is lucky enough to select a research program that achieves its goal (and does so before any rival). Previous explanations of the priority system have little to say about this facet of the system: they can only dismiss it as an unfortunate side effect of a system that works well in other respects. On my explanation, far from being a side effect, it is essential to the system's social function. By making rewards in part a matter of luck, the priority system harnesses humans' natural risk aversion so as to direct scientists to prefer more strongly than they would otherwise research programs with a high intrinsic potential for success, as required for an optimal solution to the problem of allocating resources in a winner-confers-all benefits race. The arbitrary aspect of reward distribution in the priority system, then, is the very motor of the system's being.

The Arago effect is the strict enforcement of the winner-takes-all rule, even when the interval of time between the winner's success and the runner-

up's success is very short. It has the consequence that a scientist may accrue all the prestige due to *the* discoverer by making the discovery only a matter of weeks before their competitors.

When the priority rule discriminates scientific winners and losers in a race this close, Merton writes, "priority has lost all functional significance". The Arago effect is "a dysfunctional extreme far beyond the limits of utility".[18] My analysis of the incentive structure imposed by the priority rule shows that this is not so. Discriminating between a winner and a loser in a very close scientific race, and rewarding only the winner, can benefit science and society, for the same reasons that, more generally, the priority rule is a good thing.

This point can be appreciated by comparing a fully Aragoiste priority system, which rewards only the first research program to make a discovery, regardless of how close the finish, to a more relaxed system that declares a scientific race a draw if two rival programs achieve their goals within, say, six months of one another, and rewards the participants in both programs equally. I will not present the calculations here (the comparison can be made in the same way that *Goal* and *Priority* are compared in section 4.4), but the Aragoiste scheme makes higher-potential research programs relatively more attractive, by a small amount, than does the relaxed scheme. Aragoism will therefore affect the distribution of labor among competing research programs, and so will, contrary to Merton's claim, have real functional significance.

This argument does not show, of course, that the Aragoiste system will achieve a more efficient allocation of labor than the relaxed system. Perhaps the Aragoiste system makes high-potential programs too attractive, so that the resulting distribution of labor overshoots the optimal assignment of resources to these programs. To show that the Aragoiste system is superior would involve some substantive quantitative assumptions about the degree

18. "Priorities in Scientific Discovery", p. 322.

and form of human risk aversion, and about the relation between a program's probability of its achieving its goal and its winning a priority race. (As mentioned in section 4.2, the same caveat applies to the determination of the relative merits of the *Goal* and *Priority* schemes.) My conclusion, then, is not that the Arago effect is always beneficial, but that it has a kind of effect—an effect on the distribution of labor—that may be beneficial. It is a mistake, then, simply to class the Arago effect as a pathology.

5.2 Society's Grand Reward System

I have tentatively explained the connection, asserted at the beginning of this paper, between the winner-confers-all nature of the typical scientific benefits race and science's winner-takes-all reward system. The explanation is rather complex, and the connection is therefore much less straightforward than one might suppose. Has the symmetry in payoff structure between a winner-confers-all benefits race and a winner-takes-all rewards race turned out to be a mere coincidence?

A coincidence would be a pity, both because it is more exciting to explain a striking symmetry than to dismiss it, and because the symmetry seems to have some ethical resonance for us, as exhibited in the final explanation of the priority rule in section 2.3, which makes the psychological assumption that it seems plainly fair, to us, to reward scientists precisely in proportion to the actual contribution they make to society.

In this section, I will recover the importance of the symmetry, by way of the following proposal.

1. There exists in human affairs a grand reward scheme, which determines the distribution of certain kinds of rewards, including prestige, in winner-confers-all benefits races and in many other scenarios.

2. The purpose of this grand reward scheme is to promote the optimal allocation of labor among different activities, including but not limited

to scientific research programs, whose primary goal is to contribute to the collective, not the individual, good.

3. The grand reward scheme sets a total reward for the participants in an activity in direct proportion to the actual contribution made by that activity to society. This total reward is divided among the activity's participants in proportion to their contribution to the activity's probability of success, as on the *Marge* scheme.

4. The priority scheme is nothing but the particular form that the grand reward scheme takes when applied to a winner-confers-all benefits race.

If all of this is true, then the symmetry between the benefits race and the rewards race in science is not a coincidence; rather, it is an instance of a more general symmetry that underlies the distribution of many social rewards, a symmetry mandated by the grand reward scheme.

Let me begin. To keep things simple, consider, as before, cases in which there are just two programs between which a fixed quantity of resources must be divided. Assume further that each program's goal is worth the same amount, in isolation, to society. In sections 3 and 4, I distinguished two resource allocation problems, the additive and the winner-confers-all scenarios. In the additive scenario, if both programs succeed, society receives twice as much benefit as it would if only one program had succeeded. In the winner-confers-all scenario, if both programs succeed, society receives no more benefit than it would if only one program had succeeded.

These two cases can be seen as the end points of a spectrum of distributive problems. The cases intermediate between the additive and winner-confers-all scenarios are those where, if both programs succeed, society receives more than if only one program had succeeded, but less than twice as much. An example of such a case is a scenario in which two research programs are pursuing different treatments for the same disease. Suppose

that 25% of all patients respond to both treatments, 25% to just the first treatment, and 25% to just the second treatment. Then, if just one program successfully develops the treatment it is pursuing, 50% of all patients can be cured, whereas if both programs succeed, 75% of all patients can be cured. The benefit from a second success is less than the benefit from the first success, but it is far from zero.

The resource allocation problems along the spectrum can be parametrized with a single number, which I call a problem's *additor*. The additor expresses the utility brought by a second success as a proportion of the utility brought by the first success. If v is the utility of a single success, then, and a is the additor, the utility of two successes is $(1 + a)v$. The additive case has an additor of one, the winner-confers-all case an additor of zero, and the medical example from the previous paragraph an additor of one-half.

I want to ask, first, what is the optimal distribution of labor for values of the additor partway between zero and one, and second, what kind of reward scheme might best find this distribution. Take the optimal distribution for the additive case as the baseline. In the winner-confers-all case, as I have shown, the optimal distribution favors high-potential programs considerably more than in the additive case. The optimal distributions for cases with intermediate additors fall, as one might expect, between these two extremes. They assign more resources to high-potential programs than does the optimal distribution for the additive case, but fewer resources than does the optimal distribution for the winner-confers-all case. The required bias towards high-potential programs, relative to the additive case, of course increases as the additor decreases.

How might a reward scheme implement this additor-sensitive bias? I showed in section 4.4 that the difference in bias between *Goal* and *Priority* is due to the term $s_{12}w$, where w is, in a two-program competition, the probability that the rival program wins the priority race. Call this the *priority term*. Because the priority term is larger for lower-potential programs,

and the *Priority* scheme in effect subtracts it from a program's probability of being rewarded, the effect of the priority term is to create an additional bias towards high-potential programs. Suppose that the size of the priority term were reduced by some constant proportion for all competing programs. Then the biasing effect would be smaller, because the reward probabilities after subtraction would be closer than they are in the case where the term is not reduced. One way to control the bias towards high-potential programs, then, is to implement a policy that has the effect of multiplying the priority term by $1 - a$, where a is the relevant additor. The higher the additor, the more the priority term is reduced, and so the smaller the bias towards high-potential programs.

To see how to achieve this effect, consider a program's expected total payout to its workers, which on *Goal* is just the size of the reward v multiplied by the probability of success. The effect of the priority term is to subtract $s_{12}wv$ from a program's expected payout. What is desired, then, is a policy that alters this effect so that it becomes a subtraction of $(1 - a)s_{12}wv$. The most straightforward such policy is to award the runner-up in a priority race a second prize of value av.

This means that the runner-up's payout is exactly proportional to the benefit society receives from the runner-up's success, with the constant of proportionality the same as for the winning program. In the medical scenario, for example, the first program to succeed creates a 50% cure rate; a second success increases the cure rate to 75%. Thus, the runner-up increases the cure rate by a factor of one-half, and so receives half the reward due to the winner, a proportion necessarily equal to the additor of 0.5.

I suggest that the system sketched in the last paragraph is, as I earlier intimated, close to the scheme used by society to determine the distribution of certain kinds of rewards, including prestige, for activities that benefit society as a whole. To put the proposal as simply as possible: society accords prestige and other rewards to both the winner and the runner-up of any

benefits race in proportion to the social good resulting from their respective successes.[19]

In a winner-confers-all case, where the additor is zero, the runner-up contributes nothing and receives no reward. This is, of course, the priority rule. But now it can be seen that the priority rule is an extreme case of a much more general reward scheme that matches rewards to actual contributions to society. It is with reference to the grand reward scheme that the symmetry between the winner-confers-all aspect of scientific benefits races and the winner-takes-all aspect of the priority system is explained.

Perhaps even more interestingly, our sense that the symmetry is somehow fair, even just, is, I propose, rooted in the fact that the grand scheme that mandates the symmetry is, from the point of view of the resource allocation problem, a very good one. If this is true, our sense of fairness is in some sense designed in part to solve efficiently life's many resource allocation problems.

I conclude that it is quite possible that the priority system was not purpose-built for science. It is, rather, implicit in a much more general reward scheme designed to handle the allocation of labor among projects producing any social good, whether the good is knowledge based or not. This accounts for the appearance of a concern with priority at the very dawn of modern science. The system did not have to be constructed; it was already present.

Why, then, does adherence to the priority rule seem to be a new and distinctive practice? Most social goods to which the grand reward scheme applies have additors substantially greater than zero, thus the corresponding reward schemes lack the winner-takes-all aspect of the priority system. What is new in science is the prevalence of the winner-confers-all benefits race; the

19. As I have described the grand reward scheme, it functions in the additive case identically to *Goal*. But it is *Marge*, not *Goal*, that finds the optimal allocation of resources in the additive case. There are several ways to modify the grand scheme so as to deal with the problem, but this will have to be a topic for a sequel to the present paper.

extreme values of the additors for such races give an old system dependent on the additor for its form the appearance of something new.