Herding and the Quest for Credit

Michael Strevens

To appear in *Journal of Economic Methodology*

Abstract

The system for awarding credit in science—the "priority rule"—functions, I have proposed elsewhere, to bring about something close to a socially optimal distribution of scientists among scientific research programs. If all goes well, then, potentially fruitful new ideas will be explored, unpromising ideas will be ignored, and fashionable but oversubscribed ideas will be deprived of further resources. Against this optimistic background, the present paper investigates the ways in which things might not go so well, that is, ways in which the priority rule might fail to realize its full potential as an incentive for scientists to work on the right things. Several possible causes of "herding"—an outcome in which a single research program ends up with a number of researchers well in excess of the optimum—are considered.

Herding happens when members of a scientific community overwhelmingly favor a certain paradigm, methodology, or research program, in spite of its being to science's or society's benefit that the community hedges its bets, diversifies its research portfolio, or in other words that researchers spread themselves more evenly among the available paths of inquiry. Compared to the ideal allocation of scientific labor across research programs with different goals or methods, a herding allocation results in an inefficient concentration of resources on a single approach.

Why might herding occur? You ought perhaps first to ask why it might not occur, or more generally, why anything approaching an efficient allocation of scientific labor and other resources among research programs might be expected.

To answer this question is the purpose of the first part of the present paper, in which I will present a simple model for thinking about the optimal distribution of resources among research programs (section 1), and show that the system for allocating credit for scientific discoveries and other contributions creates an incentive structure that comes remarkably close to promoting the best achievable allocation of cognitive labor (section 2). In the second part of the paper (section 3), some reasons for deviation from the optimum, particularly those resulting in herding, are considered against this backdrop.

1. Optimality Ideally

In 1876, Charles Peirce addressed the problem of efficiently allocating resources among research projects in an addendum to the annual report of the United States Coast Survey (Peirce 1879). He imagined two or more research programs, each yielding a certain quantity of knowledge for a given investment of time and other resources, and asked how to distribute a fixed quantity

of resources among the programs so as to maximize the return in knowledge.

Peirce transforms this task into a straightforward mathematical calculation by way of three assumptions. The first is the assumption of a fixed, determinate function—the *returns function*—relating resources invested to knowledge produced. Implicit in this assumption is the irrelevance of contextual factors: only direct investment influences epistemic output.

The second is the assumption that the returns function is strictly increasing—so investment of more resources always brings some return—but that marginal returns are decreasing, so that the return for investing a fixed quantity of resources in a program grows smaller, the greater the amount of resources already invested in that program.

The third and final assumption is that the total value of the knowledge returned by several research programs is additive, that is, is equal to the sum of the value of the knowledge returned by each. This assumption does not hold if, for example, two programs compete to make the same discovery, since the value of uncovering the same fact twice is not double the value of uncovering it once. As I will show in the next section, this assumption can be weakened with interesting consequences, but for now let me stay with Peirce's original model.

There are some other assumptions implicit in the very question that Peirce asks, for example, that there is a single scale of value, of utility, against which the returns from any assignment of resources among programs can be measured, or that scientific inquiry can be divided into distinct, separately funded research programs. I will accept these assumptions for the most part without further comment.

Peirce's assumptions characterize a simple model of the problem of resource allocation. For each research program there is a returns function with decreasing but always positive marginal returns; the problem is to determine the allocation of resources that maximizes the sum of the functions. Focusing (as does Peirce, for expository purposes) on the case where there are two

programs only, the returns functions may be written $R_1(n)$ and $R_2(n)$; the aim is then for a given total quantity of resources N to find the value of n that maximizes $R_1(n) + R_2(N - n)$. When N is large enough, the maximum allocation will be found where the derivatives of the two functions are equal, that is, at the point where one additional unit of resources would bring an equal marginal return if invested in either program. (For simplicity's sake, let me for the most part ignore, in what follows, the case where the maximum requires that all resources be directed to just one of the two programs.)

As Peirce notes, the maximum may be found by following a hill-climbing algorithm. If you imagine the fixed quantity of resources becoming available one unit at a time, then you can ensure that they are optimally distributed by sending each latest increment to whichever of the two programs offers the greater marginal return for the investment. The same is true if the total quantity of resources to invest is not fixed but increases forever: always allocate incoming resources to the program that yields the greater marginal return, and you are guaranteed that at every moment from now until eternity, you have achieved the best possible distribution of the available resources. This holds regardless of the number of research programs clamoring for money, brains, and equipment.

To implement the Peircean hill-climbing method, you ought ideally to be all-knowing, all-powerful, and benevolent. *All-knowing*, you grasp the form of the true returns functions, so that you can correctly determine which investment will bring the greatest marginal return. *All-powerful*, you have complete control of the available resources, and so can effect whatever distribution seems best to you. *Benevolent*, the distribution that seems best to you is the distribution that is best for society as a whole, presumably maximizing the quantity of knowledge produced, weighted by the social and intellectual value of that knowledge.

Peirce's paper was written for his superior, the Superintendent of the Coastal Survey, who he might reasonably have supposed to approximate

these qualities—all-powerful for bureaucratic reasons, all-knowing insofar as much of the research undertaken by the Survey was applied science in which the returns for a given investment were reasonably well understood, and socially benevolent in that nineteenth-century fashion typical of the aristocracy to which the superintendents belonged (as did Peirce's own father, superintendent from 1867 to 1874).

In modern research science there are no more superintendents. But let me imagine a Superintendent existing in ideality if not reality. This fictional person is all-knowing, I stipulate, in the sense that they are in possession of all information currently available to science; they may not know the "true" returns function for any given scientific research program, but they know as much as there is, presently, to be known. Their best guess as to the form of the function, then, is as good as anyone's guess. Consequently, they represent an epistemically relative ideal, the closest thing to the true returns function to which we in principle have access, given the current state of scientific knowledge.[1]

The Superintendent is benevolent in precisely the nineteenth-century fashion: they seek to maximize the total social benefit, in the broadest sense, returned by scientific research. And they are all-powerful, in that they have total control over the distribution of scientific resources. Most important—since my ultimate topic is herding, which is in the first instance a matter of the distribution of scientists rather than of other resources—they have total control over which scientists work on which research programs.

The Superintendent's notional decisions as to how to distribute cognitive labor may be understood as a kind of ideal against which actual allocations of resources in science can be measured. Such an ideal is not, of course, beyond criticism. Now that we no longer live in the nineteenth century, the notion of

1. From a Bayesian perspective, the Superintendent will not have a single "best guess" but a probability distribution over possible returns functions. In what follows, talk of a "best guess" stands in for all such other implementations of the Superintendent, on the details of which nothing in this paper turns.

a unified measure of social good is somewhat suspect (even if such measures are frequently used to construct and evaluate public policy). Likewise, it is unclear that the diverse views of the world's many scientists can be melded into a single "best guess" at the true returns function. But put these worries aside; this is a paper about herding in a volume about herding, and the very notion of herding presupposes a fact of the matter about the ideal allocation of cognitive labor, from which when herding occurs, the actual allocation departs. The notional creation of a Superintendent may underestimate the problems in constructing such an ideal, but a first approximation to the ideal will be enough to motivate some potential explanations for herding.

In the same spirit, I will throughout this paper retain a number of Peirce's other assumptions, while recognizing that at best they hold only approximately true: that returns functions have no externalities, so that a research program's return is determined entirely by the amount of resources directly invested in that program; that returns functions offer strictly decreasing marginal returns; and so on. (The additivity assumption will, however, be relaxed, as noted above.) The aim is to find out whether, even under the relatively ideal conditions represented by these assumptions, there are circumstances in which herding is liable to occur.

Putting these principled difficulties aside, what practical impediments are there to real science's mirroring, in its allocation of resources and of cognitive labor in particular, the Superintendent's ideal allocation? There are two.

First: except under extraordinary circumstances (immanent climatic catastrophe, for example), there is not the kind of amalgamation of information that allows all relevant parties to come to know the profession's "best guess" as to the form of the true returns function. Indeed, there are incentives for scientists competing to realize some research goal to withhold information about the prospects of their programs from one another. Still, you might suppose that broadly correct estimates are possible (where "broadly correct" means "close to the Superintendent's best guess"), and so that ignorance does

not stand in the way of a broadly optimal allocation of resources.

The second impediment is far greater: decisions as to which scientists devote their time and energy to what projects are typically not made by a benevolent Superintendent, but by the scientists themselves. It is scientists, or putative scientists, who decide what to study at graduate school, what post-doctoral positions to pursue, and how to set up their labs. To some extent, grant-giving bodies can play the role of the Superintendent. But while these sources of largesse dispose, they do not propose: they are essentially reactive, and so provide at most a corrective to the allocations of labor effected by the personal decisions of individual scientists.[2]

If each researcher thought like the Superintendent, surveying the opportunities for social amelioration offered by the scientific enterprise and meekly going wherever the need was greatest, then perhaps this difficulty would not be so substantial. But there is good reason to think that scientists are rather differently motivated, as Peirce himself noted in the cautionary final paragraph of his resource allocation paper:

> It is to be remarked that the theory here given rests on the supposition that the object of the investigation is the ascertainment of truth. When an investigation is made for the purpose of attaining personal distinction, the economics of the problem are entirely different.

2. Optimality in Reality

How does a scientist decide to specialize in a certain domain, to inhabit a particular intellectual niche, to devote their days to one particular problem

2. I exaggerate to some degree; for example, the NSF and other grant agencies sponsor initiatives to draw scientists to work in areas considered undersubscribed, and grant money determines where there are post-doctoral positions to pursue in the first place. The power of the NSF to determine the distribution of cognitive labor is nevertheless minuscule compared to that of the Superintendent.

above all others?

Interest and personal history play a role, but it has long been supposed by sociologists of science that the most powerful force in such decisions, for amateur and academic scientists, is the quest for credit (Merton 1957; Latour and Woolgar 1986).

What is credit? It is the standing that a researcher accrues in virtue of their past contributions to knowledge and know-how—their standing, in the first instance, in the scientific community rather than the world at large. Credit is, then, what other writers have called reputation, credibility, prestige, status, fame. (Perhaps these are not exactly the same thing, but fine distinctions will not matter in what follows; I will proceed with a perhaps unrealistically monolithic conception of the social rewards of good science.)

What is the value of credit? There is on the one hand an intrinsic satisfaction to be found in the esteem of other scientists, well captured by Paul Samuelson's famous words:

> In the long run, the economic scholar works for the only coin worth having—our own applause. (Samuelson 1962, 18)

On the other hand, this same esteem is of the utmost instrumental importance: from prestige flows the resources to do more science, namely, grant money, lab space, and graduate students, to say nothing of a larger office and a higher salary (as emphasized by Latour and Woolgar 1986). In short, this intangible social resource *credit* provides rewards intellectual, emotional, and material, as well as the wherewithal to go on doing science at a similarly high level, thereby generating yet more credit.

The prospect of credit is therefore a powerful motivator in any research-related decision a scientist might make, not least the decision as to what research program to join or to instigate in the first place. It is not the only such motivator. There are other intellectual and moral motivators, such as curiosity and the desire to increase human happiness. There are other material

motivators, such as the prospect of lucrative patents or positions in industry or government. In what follows, however, let me focus on credit.

The Peircean Superintendent distributes resources among research programs so as to maximize overall social good. If the Superintendent is wise and competent, society benefits accordingly—so Peirce promises. But if the prevailing sociological wisdom is correct, then in real science, or at least in real academic science, what drives decisions about the allocation of cognitive labor are not the aloof and benevolent intentions of a central planner, but rather individual scientists' independent strivings for credit. Investigations are undertaken, in Peirce's words, "for the purpose of attaining personal distinction". What hope, then, of a socially satisfactory outcome?

To answer this question—far more optimistically than Peirce himself anticipates—let me begin with the ideal rather than the real. Under what circumstances will credit-seeking scientists distribute themselves among research programs in just the way that the notional Superintendent would distribute them? In what circumstances, that is, will scientists in pursuit of their own private ends achieve a socially optimal distribution of labor? (Or rather, a distribution as socially optimal as the present state of knowledge, epitomized by the Superintendent's best guess at the relevant returns functions, allows?)

What circumstances make a difference to the distribution of credit-seeking scientists' labor? Above all, the social norms for distributing credit itself. In principle, such norms might take any form you like. All scientists might be accorded equal credit, regardless of their contributions to the store of scientific knowledge. Scientists might be accorded credit in proportion to the energy and time that they put into their researches, regardless of the outcome. Or—closer to the scheme observed by sociologists in real science—credit might be proportioned to actual achievement, regardless of the time, talent, energy, or other sacrifice required to produce that achievement. Possible norms differ also in the way they measure the value of an achievement. Some may take

into account only the practical value of a discovery. Some may take into account only a discovery's theoretical value. In the latter case, some may take into account the satisfaction that the general populace takes from scientific knowledge, while others may care only for the intellectual stimulation of specialists. Any of these choices seems possible; indeed, each of them is familiar as a scheme for distributing something or other of worth in human society.

You might, then, imagine a lazy variation on the Peircean Superintendent who, rather than doing the hard work of allocating cognitive labor themselves, decides to use their unlimited power over the world of science to determine in advance a rule by which credit will be doled out, selecting a scheme that will ensure that scientists make decisions about their own future research that effect the very same distribution of labor that a more assiduous Superintendent would have commanded directly. My questions: Is such a scheme available? If so, is it the kind of scheme that would be humanly acceptable—does it comport well enough with basic human notions of fairness and reasonableness? And how close is it to the actual scheme for rewarding researchers with credit?

In answering the first of these questions—in determining what rules for distributing credit will amount to an incentive scheme that, as a byproduct of individual scientists' self-interested credit-seeking behavior, brings about a socially optimal allocation of labor—I will make several simplifying assumptions. First, I continue to focus on the allocation of cognitive labor and not at all on the distribution of other resources necessary for doing science, such as expensive experimental equipment. Since my topic is herding, and herding is a matter of the distribution of minds, this is not a significant distortion. Second, I assume that a scientist's decision as to which research program to grace with their intellect is entirely determined by the desire to maximize the amount of credit earned. The ultimate question is whether credit-seeking behavior distorts such decisions for better or worse; the effects of other motivators will be put aside. Third, I make the same assumptions that Peirce makes: a pro-

gram's returns are entirely determined by the quantity of resources invested; its marginal returns are decreasing but non-disappearing, and there is a univocal scale of social good against which any particular basket of returns can be measured. I also at first assume that, according to the scale of social good, the aggregate social value provided by any two research programs is equal to the sum of the values of each in isolation—the assumption of additivity.

Suppose that several research programs are competing for resources. Peirce's Superintendent, you will recall, is able to maintain the optimal allocation of resources at all times by following the hill-climbing rule: assign the next unit of resources—the next scientist-hour—to whichever program offers the higher marginal return upon the investment of that resource. Credit-seeking scientists will mirror the decisions of the Superintendent if they distribute themselves so as to conform to the same rule. They are not, of course, following the rule consciously; rather, they choose to join, or to devote their next week or month or year of thinking time to, the program that offers the higher return, in terms of credit, to them and them alone. But the behavior of the scientists and the notional decisions of the Superintendent will coincide if the program that offers the highest marginal return to society, for the scientist's investment of labor, also offers the highest return in the form of credit to the scientist.

There is an obvious way to ensure that this is so: simply reward scientists in proportion to the marginal social return generated by their labor. For committing a certain amount of time, energy, and talent to a research program, then, a scientist should receive credit in proportion to the resulting increase in the program's social returns, or in other words, they should be rewarded in proportion to their contribution to the social good. Implement such a norm for the allocation of credit, and scientific labor will distribute itself so as to maximize social good (Strevens 2003).

Two remarks. First, this result holds true for any measure of social good, whether theoretical, practical, or a mix of the two. I will avoid having to

pronounce on this question any further by adopting the Superintendent's definition of the good.

Second, I have assumed so far—as does Peirce—that the return from a research program, for a given level of investment, is guaranteed. In reality, research programs sometimes fail to realize their goals, or they are scooped, meaning that some rival program reaches the goal first, thereby rendering the efforts of all other programs toward the same goal otiose. Peirce's model is easily applied to science in all its stochasticity, however, by understanding the returns function as representing not the actual but the expected social value generated by the program for a given level of investment. Credit-seeking scientists will in that case distribute themselves so as to maximize social good if they are rewarded for investing in a program in proportion to their expected contribution to the program's returns. No formal adjustments to the model are required in order to represent the risks inherent in empirical investigation, then; an adjustment to the model's interpretation is enough.

Call the norm for allocating credit suggested above the *expected contribution* rule. How feasible is the rule? How realistic? Quite realistic, and therefore quite feasible, is the answer: scientists do seem to earn credit in rough proportion to the social payoff that can be attributed to their efforts, both in a narrow intellectual sense and a broad practical sense, which is to say that they are rewarded both for plumbing the deep structure of the universe and for discoveries that save and improve lives, the more so the more lives are saved or bettered. In short, the actual rule for allocating credit in science seems rather well tuned to the Superintendent's wavelength, resulting in a pattern of decision-making by individual researchers that tends to find the most socially beneficial distribution of cognitive labor among research programs.

A closer inspection shows, however, that the actual rule of credit allocation differs in two crucial, related respects from the expected contribution rule. First, the actual rule is not an *expected* contribution rule but an *actual* contribution rule: scientists receive little or no credit unless their research

program succeeds in its goals. With success comes credit in proportion to a scientist's contribution to the probability of that success, then, but with failure comes nothing at all, no matter how greatly a scientist's efforts increased the probability of success.

Second, success itself is not sufficient for credit. A research program must not merely make a contribution to knowledge; it must be the first to do so. It is for this reason that the reward system in science is conventionally referred to as the *priority rule*. When, as is often the case, two or more scientists vie with each other (knowingly or not) to make the same discovery—the development of calculus, the observation of the planet Neptune, the structure of DNA, the cause of AIDS—and when they both independently succeed, only the first to do so is held to have any claim to the credit for the achievement. Thus the many heated disputes over who of two nearly simultaneous discoverers has the claim to priority: Leibniz versus Newton, Le Verrier versus Adams, Montagnier versus Gallo (Merton 1957).

Scientific credit is not assigned according to the expected contribution rule, then, but according to the priority rule, which is a species of actual contribution rule. Does it matter? Will scientists' decisions be any different than under the expected contribution rule? As I will explain shortly, decisions made under the priority regime will deviate systematically, if subtly, from the Superintendent's diktat in an interesting way. But for now let me turn to an apparently unrelated topic, Peirce's questionable assumption of additivity.

Returns from two or more research programs are additive if the social value of the aggregate returns is equal to the sum of the values of the returns of each of the programs in isolation. The additivity assumption would seem to hold for programs that offer benefits satisfying independent needs, for example, understanding speciation and preventing malaria. But it begins to fail when programs promise alternative ways of satisfying the same need, and it fails dramatically when programs have the very same goal, that is, when rival programs attempt to make the same discovery—as exemplified above by

the race to find the cause of AIDS, among other historical cases.

The reason is this: once a discovery has been made, a further iteration of that discovery is worthless. Gaining the same knowledge many times over is no more beneficial than gaining it once (except insofar as the second and subsequent discoveries may generate novel collateral results). The social value of many instances of the same discovery is not the sum of the values of each discovery in isolation; it is nothing over and above the value of a single such discovery.

Let me focus on this case of radical additivity failure, where two or more programs pursue the same end.[3] For expository purposes, let me make some further simplifying assumptions: there are two competing research programs; each program either succeeds or fails completely, yielding either the full social benefit it is capable of providing or none at all; and the benefit promised by each program is the same (as is the case when the programs pursue the same result).

From the Superintendent's point of view, the problem may be stated as follows. When additivity holds, the optimal social benefit comes from the allocation of resources that maximizes the expected number of successes (since I am assuming that the successes of each program are equally valuable). When additivity fails because one success renders other successes valueless, the optimal social benefit comes from the allocation of resources that maximizes the probability that at least one program succeeds. Formally, in the additive case what should be maximized when there are two programs is

$$p_1(n) + p_2(N - n),$$

where $p_1(n)$ and $p_2(n)$ are the probabilities of each program's success given the investment of n resources and N quantifies the total resources to be

3. In other cases additivity fails less completely; for example, two cures for the same disease may, if the cases where they do some good overlap but not entirely, provide more aggregate social good than either cure in isolation, but not so much as the sum of the goods provided by each in isolation.

distributed. In the extreme non-additive case, the race to make the same discovery, what should be maximized is

$$p_1(n) + p_2(N - n) - p_{12}(n, N - n)$$

where $p_{12}(n, m)$ is the probability that both programs succeed, given the allocation of n resources to the first and m to the second. If $p_{12}(n, m)$ is in general non-zero (if the two programs' successes are not mutually exclusive), then the solutions to the two problems will differ: the best allocation of resources for the additive case will be less than optimal for the non-additive case.

To better appreciate the difference between the optimal distributions for the additive and non-additive cases, suppose that one of the two research programs has strictly *higher potential* than the other, in the sense that the probability of the one program's reaching its goal, given a certain investment, is always greater than the probability of the other program's reaching its goal given the same investment. In symbols, the first program has a strictly higher potential than the second just in case $p_1(n) > p_2(n)$, for any choice of n (figure 1). In the additive case, the optimal investment of resources always allocates more to the higher- than to the lower-potential program. In the non-additive case, the optimal investment favors the higher-potential program even more strongly. A Superintendent following the hill-climbing rule, then—always allocating resources to the program that yields the higher marginal return from the investment—will underinvest in the higher potential program. For the same reason, credit-seeking scientists rewarded according to an expected contribution scheme will distribute themselves so that too few are working on the higher-potential program.

Is there a reward scheme that will do better than the expected contribution scheme in the non-additive case? As Strevens (2003) shows, there is: namely, an actual contribution scheme, and in particular the priority rule. (I remind you that in the present scenario, in which programs are racing to make the same discovery, the priority rule rewards scientists in proportion to their

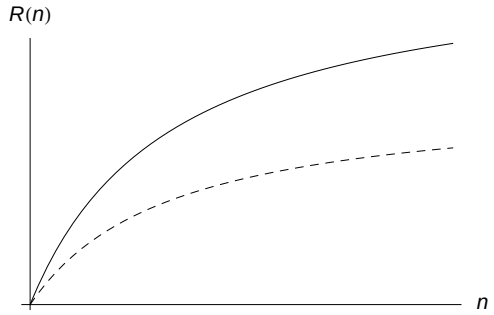

Figure 1: Returns functions for higher-potential (solid) and lower-potential (dashed) research programs. For any investment of n resources in both, the expected return $R(n)$ from the higher-potential program is greater than the expected return from the lower-potential program. When programs pursue the same goal, this is equivalent to the probability of success $p(n)$ being greater, for all values of n, for the higher- than for the lower-potential program.

expected contribution to the program they join, but only if the program actually succeeds in making the discovery in question and is the first to do so.)

The priority rule is better in a non-additive race because, although it creates distributions of labor that broadly resemble those created by the expected contribution rule, it systematically favors higher-potential programs more than does the expected contribution scheme. Under a priority regime, then, a few more scientists will tend to work on higher-potential programs than under an expected contribution regime; the priority rule will come closer, then, to achieving an optimal distribution of resources among programs that compete non-additively.

Why does the priority rule create an additional bias toward higher-potential programs? Two reasons. First, under the priority rule, scientists are rewarded only if their program succeeds. The expected contribution rule takes into account the size of the probability (which of course determines, along with

the value of the goal, the expected contribution), but it does not take into account risk aversion. Risk-averse scientists will, given the choice between two programs to which they would make the same expected contribution, prefer to join the program that is more likely to succeed and therefore bring any reward at all. This is the higher-potential program; hence, under an actual contribution regime scientists will favor higher-potential programs more than they do under an expected contribution regime.

Second, under the priority rule, the success of a program is not sufficient for the program's scientists to be rewarded: the program must also succeed earlier than any competing program with the same goal. If it is assumed that higher-potential programs will tend to finish earlier than lower-potential programs when they do succeed—if it is assumed that the same properties that make for a higher success probability make for a shorter expected waiting time to success—then the crucial role of priority in the allocation of credit will motivate scientists to favor higher-potential programs more still. (The assumption of a correlation between waiting time to success and probability of success is quite empirical, but also I think quite reasonable.)

A small miracle is in evidence. It seems that the norm for distributing credit in science is finely calibrated to lead credit-seeking scientists to distribute themselves among research programs so as to maximize the overall social good those programs stand to bring, as though the Superintendent had come down among the scientists in some formative period and arranged things just so. Even the apparently rather brutal priority rule, which denies any credit whatsoever to a scientist who makes a discovery a week too late (Merton 1957, 658), turns out to be aligned to a higher good, adjusting scientists' behavior as it does to better find a social optimum in non-additive competitions.[4]

4. Some connections to the recent literature: Kitcher (1990), without discussing the priority rule explicitly, argues that self-interested scientists may distribute themselves among research programs in a way that is more advantageous to society as a whole than the distribution achieved by high-minded scientists. Dasgupta and David (1994) propose other functions

There are limits to this functional paradise. The effect of the priority rule is presumably not precisely calibrated to the mathematics of non-additivity, depending as it does on independent variables such as the average level of risk aversion and the correlation between a program's chances of success and its waiting time to that success. Thus although the changes to the expected contribution rule that transform it into the priority rule nudge the resulting distribution of labor in the direction of optimality, there is no reason to think that they find the exactly maximal distribution.

More important, the optimum that is approximately attained is optimal relative to the Superintendent's best guess as to the true returns functions, not to the functions themselves. It is relative, that is, to what the returns functions would seem to be if all current knowledge were taken into account—but current knowledge might be misleading. Further, scientists can hardly be expected to have at their fingertips even all the available information; thus they will fall short to some extent of the Superintendent's best guess, which itself falls short to some extent of the truth.

Finally, scientists are presumably not pure credit-seekers; sheer curiosity and perhaps a desire to help humanity will play some role in their decisions as to where to send their brainpower, and so these decisions will depart to some extent from the choices urged by the priority rule.

For all that, the priority rule finds a pretty good solution, under the circumstances, to an extremely difficult coordination problem. Think of the ways in which the human race usually copes with such problems: you will see either sporadic chaos (stampedes, exhausted fields and fisheries, financial collapses, and so on) or rigorous, computationally intensive central planning. Somehow the scientific reward system avoids both. A simple rule guides scientists to something that is at least creditably close to an optimal distribution

for the priority rule to play in science, such as the provision of motivation to publish as soon as possible. For some reservations about economic approaches to thinking about the social structure of science in general, see Hands (1997).

of cognitive labor, requiring of them only some degree of savvy about the prospects of the programs in their field and a certain level of self-interest. It is as though an invisible Superintendent is at their side, directing them to wherever they are most needed. But of course the Superintendent is not invisible at all. It is the reward system of science.

3. Herding

The point of this paper is not to praise the scientific reward system, and the surprisingly satisfactory distribution of intellectual labor that it appears to incentivize, but to ask how it might go wrong, and in particular, how it might allow, or even encourage, herding—how it might result in far more scientists joining a research program than the Superintendent would advise.

Under what circumstances will scientists overly favor a single research program? There are various malfunctions of the scientific reward system that might lead to herding:

1. Scientists might deliberately abandon the priority rule—if, for example, it were perceived to be unfair and in need of reform.

2. Scientists might for some reason lose their taste for credit, for status, for reputation—hard though that might be to imagine.

3. Other incentives might swamp or otherwise distort the motivating force of credit. Then, although single-mindedly credit-seeking scientists would distribute themselves optimally, fully human scientists might not—if, for example, they received enormous grants from politically or economically motivated actors to join a particular research program regardless of its meager prospects.

4. Scientists might badly overestimate or underestimate their expected contribution to a research program, overshooting by far the Superintendent's best guess.

The latter two are more plausible explanations of present-day herding, I think; in the remainder of this paper I will focus on the last, asking in particular how scientists might systematically overestimate the contribution they can reasonably expect to make from joining a certain research program, resulting in an inefficiently large concentration of cognitive labor on that program.

I will continue to suppose, for expository purposes, that some research programs have higher potential than others, where for any given investment of resources, a higher potential program yields a higher return than a lower potential program (see figure 1).

Consider a series of scientists choosing between two as yet unstaffed programs, one with higher potential than the other. The first such scientist will choose the higher-potential program. What about the second scientist? Their expected contribution to the higher-potential program will be somewhat less than the first scientist's, because of the returns function's decreasing slope. As more scientists join the higher-potential program, the marginal return, and thus each next scientist's prospective contribution, continues to decrease, until it drops below that offered by the lower-potential program. When things go according to the Superintendent's plan, then, it is the decreasing marginal returns that prevent a program's becoming overstaffed: the more scientists who join a program, the less attractive that program looks to the next scientist.

This mechanism will function as specified, putting a brake on scientists' joining the higher-potential program, provided that everyone's opinion of the program's underlying potential—its returns function—stays constant. But what if their opinion of the underlying potential improves as the number of scientists working on the program increases? What if this improving opinion makes up for, or more than makes up for, the braking effect? Then as more scientists join the program, its attractiveness to new credit-seeking recruits will not decrease, and may even increase. As long as this goes on, scientists will join only the one program, ignoring all others.

Is this herding? It depends on the reason for the increase in the opinion of

the program's underlying potential. Let me discuss two possible mechanisms for the increase. First, it may be that scientists come to see the program as having more potential because, as more scientists work on the program, they reveal underlying strengths in the program that were not previously known. The improvement in opinions is due, in this case, to improvements in knowledge about aspects of the world on which the success of the program depends. Second, it may be that scientists do not learn anything directly about the world, but that they react to other scientists' joining the program by supposing that there must be some relevant information about the world of which the joiners, but not the observer, are aware. Without knowing what this information is, they nevertheless revise their opinion of the program upward.

Consider the first of these mechanisms, which is driven by the acquisition of new information about the world. Let me take a moment to discuss the way in which the reward system deals with this epistemic dynamic. In Peirce's setup, the Superintendent's best guess as to the returns functions never changes. For a given quantity of resources, then, the optimal allocation will not change. In real science, however, new information inevitably will come to light, in view of which the Superintendent's informed opinion about the returns functions will change. Thus, what seems optimal will evolve, or if you like, what was optimal relative to the old epistemic situation will be suboptimal relative to the new situation. If suboptimally distributed resources can be reassigned to implement the new optimum, the Superintendent will make it happen. In the case of cognitive labor, however, the resource is hours spent, something that cannot be retrieved without turning back time. The Superintendent is all-powerful, but not *that* all-powerful. The best they can do is to use new resources, as they become available, to repair the suboptimality, working toward the optimal allocation for the new quantity of resources.

The Superintendent is fictional, of course, but the reward system in science reacts to new information in much the same way. As new information appears, scientists' estimates of the expected contribution they will make to the various

available programs will change, and so their opinion as to which program will offer the most enticing return, in terms of credit, will change along with it. As a consequence, they will, unwittingly, do just what the Superintendent would do: they will distribute their time so as to proceed from a distribution newly revealed as suboptimal to the optimal distribution as fast as is feasible. By mirroring the Superintendent's actions as closely as possible, then, the reward system implements the best distributional policy that is possible given the available information about programs' true returns functions.

Herding means not just scientists' flocking to a single research program, but also their doing so suboptimally. In the case where new research continually turns up evidence pointing to greater and greater underlying potential for the program in question, what the scientists do is not suboptimal at all. Thus it is not herding, but a reasonable and desirable recalibration of the distribution of resources. Sending all those scientists to a single program is just what the Superintendent would do, and without knowing more than the Superintendent, that is the best that can be done.

Consider now the second mechanism by which, as more scientists join a program, other scientists' estimates of the program's underlying potential may improve. In this case, the scientists revising their opinions are not learning anything about the world. They presume that the scientists joining the program know more than they do, and it is this putative additional information that improves their opinion of the program's underlying potential. But there is in fact no such information; the scientists joining the program do not know any more than the rest. (Or let me stipulate that they know no more, for simplicity's sake.)

The upward revision of opinion is an error, then, in the sense that it does not reflect a change in the views of the Superintendent, who has all information available to the scientific community at their fingertips. From the Superintendent's perspective, the influx of scientists to the relevant program is suboptimal, based as it is on information that in fact reveals nothing new

about the program's prospects. It is a case of herding.

The cause of the herding is what might be called an *epistemic mirage*, in which scientists infer the existence of evidence that no one in fact possesses. Epistemic mirages are mild versions of what are now called *information cascades*. In an information cascade, epistemic feedback entirely overwhelms genuine information (Bikhchandani et al. 1992). For herding to occur, nothing so drastic is needed. The mirage need only put a finger on the scale, adding illicitly to genuine information without necessarily rendering it irrelevant.

Under what conditions will herding due to an epistemic mirage occur? A necessary condition is that individual scientists are not, and know they are not, in possession of all available information. This is what allows observers to wonder, of scientists joining a research program, "Do they know something I don't know?".

Even if they give this question a positive answer—even if they conclude that the joiners do have information that they, the observers, lack—the mirage need not cause herding. The observers might quite reasonably conclude, after all, that they have information that the joiners lack. For herding to occur the observers must conclude that the joiners have information that is not only different, but better. Only then will an epistemic mirage tilt the decision-making dynamic in favor of the joined program. But why should the observers think that the joiners have, on average, better information? Why should they think that the joiners are better informed than the average scientist, or that they, the observers, are less well informed than average?

In some cases where epistemic mirages are possible (including many of the usual examples of information cascades), the fact that the one party has taken decisive action—walking into the restaurant or selling the securities—gives the other, observing party some reason to think that they, still hanging back, have less information. But scientists typically do not have the luxury of hanging back. They must go to work on something or other. So they have no reason to think that joiners are better informed than the rest. There is no

"rest"—everyone is, of necessity, a joiner.

Indeed, the observers who matter for the purposes of understanding herding in science are about to become joiners. They have just finished their doctoral degrees, or their post-docs, or their present research project is about to draw to a close. What reason do such scientists have for relative epistemic insecurity? They are slightly junior, on average, to the joiners, but since every observer is about to be a joiner, and every joiner was the moment before joining an observer, the effect of seniority is surely vanishingly small.

It is hard to see, then, how epistemic mirages will cause herding under the conditions in which scientists decide where to commit their labor. Or at least, this is so provided that scientists are at least somewhat reflective and not prone to irrational bouts of epistemic insecurity.

I have considered two mechanisms that might cause herding. The first is quite real, but what it causes turns out not to be herding. The second causes herding, but there are serious doubts as to whether it is much seen in science. Let me therefore discuss a third mechanism.

When a scientist's research program succeeds, and is the first to do so, the scientist is rewarded according to their expected contribution to the program. Who determines the expected contribution? In the Peircean paradise, it would be the Superintendent: their social utility function would determine the size of the pot (equal to the total social contribution of the program), and their best guess as to the returns function would determine the scientist's personal share of the pot (proportional to their marginal contribution to the probability of success). In the real world, these determinations are made by scientists themselves, since one scientist's reputation subsists in the attitudes of all the others. (In cases of life-saving or world-historical achievement, society at large may play a role, but most scientists cannot realistically expect this degree of success.) A narrowly credit-seeking scientist ought, then, to choose not the program to which they expect to make the greatest contribution, but the program to which they expect other scientists will regard them as having

made the greatest contribution. The reward system in science is a Keynesian beauty contest.

Insofar as all scientists think alike, this fact has no significant consequences; the scientist might as well use their own expectations as proxy for the expectations of scientists generally. In certain circumstances, however, the two might diverge.

Consider a thoughtful credit-seeking scientist, wondering whether they should join a research program that has as yet little recognition. In addition to all the usual considerations (including risk aversion and the importance of speed in a priority race), the scientist faces the following Keynesian worry: what if, by joining the program, they make a significant contribution that is not recognized as such? Apart from the ghostly approbation of the Superintendent, they will gain nothing for their success, at least in the medium term on which their survival as a scientist may depend.

To the extent that such worries are reasonable (I continue to assume that my scientists are fairly rational), there exists the potential for an unwarranted bias—a departure from the Superintendent's allocation—in favor of large, established research programs.

When, then, might scientists reasonably worry that their contribution to a smaller research program would not be appreciated for what it is? Most notably, when the contribution of the research program itself will not be adequately appreciated.

In many circumstances, a scientific achievement will push its way into the scientific spotlight on its own merits. The successful prediction of the degree to which light is affected by gravity, or of the degree of red shift in receding galaxies, or of the behavior of the planet Mercury's orbit—these are accomplishments that will not be ignored by reasonable people. They revolve around the derivation, from a theory, of precise numerical predictions (or retrodictions) whose match to reality cannot but impress informed observers.

In some domains of science, however, precise forecasts or retrocasts are

rare, or at least are rarely correct. Achievements in these domains revolve around fitting phenomena into an explanatory framework.

Sometimes this framework takes the form of a mathematical model with adjustable parameters. The model cannot be said to predict or retrodict the phenomena in question, because the model's parameters are tuned "by hand" to yield the correct result. The point of fitting the phenomena to the model is to show that a certain kind of explanation is a plausible one. Evolutionary theory, for example, often works with mathematical models that reproduce the direction or trend of evolution with some precision, but with relatively little predictive power because the values of parameters such as selection coefficients are derived in an essentially ad hoc way from the data itself.

Sometimes an explanatory framework is not quantitative at all. Much explanation in the social sciences is like this: there are many explanations but relatively few mathematical models in anthropology for example.

The value of fitting data to explanatory frameworks in this way can be very great; the rise of evolutionary theory is built on just such achievements. A well-tailored explanation, then, can make a considerable contribution to the advancement of science. When a theory explains a datum by its own lights, however, the achievement may not be so obvious to the outside world. So perhaps research program *A* can fit many phenomena to research program *A*'s explanatory framework . . . but what does that matter if program *A* is unlikely to pan out in the long term? If scientists think along these lines—and it is arguably quite reasonable for them to do so—then the non-predictive explanatory achievements of lower-potential theories will be systematically underestimated in the short and medium term. But then scientists will be reluctant to join such programs, in case they fail to receive credit in proportion to their theoretical contribution in a time window that is crucial to professional survival.

Does their reluctance stand to create a suboptimal distribution of resources? A bias in favor of higher-potential programs is not necessarily a

bad thing; such a bias is, after all, what makes the priority rule superior to the expected contribution rule for allocating labor in non-additive scenarios. But in the present case, the degree of bias depends greatly on a factor that has no influence on what distribution of labor is optimal, namely, whether a research program is more of the predictive or more of the explanatory kind. Of two programs with equal potential—with identical returns functions—the bias will more strongly dissuade scientists from joining the more explanatory program. That cannot be optimal.

My next question: can the bias operate so as to favor a single program far more strongly than the Superintendent would favor it, thus causing herding?

I think that the answer is yes. Let me tell a story in two stages. In the first stage, a budding science offers a number of research programs, or better, a number of kinds of research program. One of these kinds is predictive; the others are explanatory. (This does not mean that the one makes good predictions and the others bad predictions; the one's predictions might be quite inaccurate.) Suppose that the various programs have roughly equal potential. Choosing among them, scientists will nevertheless—if the bias I have been describing is real—show a preference for the predictive program. The preference is perhaps mild, so they will not shun the explanatory programs, but they will join them in smaller numbers, taking into account the possibility that whatever contribution they make to them will be more difficult to establish to the field's satisfaction in the short to medium term.

In the second stage, a certain kind of feedback operates. As the relative size of the predictive program grows, scientists quite reasonably see that a smaller and smaller number of their colleagues, in relative terms, are in a position to appreciate explanatory contributions to a primarily explanatory program. (Note that as in an election, it is the proportion of appreciators, not the absolute number, that determines the amount of credit rewarded.) Thus their worries grow in seriousness as the predictive program grows in relative size. The bias's effects enhance the strength of the bias.

There is a natural brake on this process: as more scientists join the predictive program, its marginal returns and so the joiners' expected contributions decrease, making the program ever less attractive. But the competitor programs—the explanatory programs—also become less attractive at the same time, because of the feedback. If the latter effect outpaces the former, then relatively ever more scientists will join the predictive program. Herding will happen.[5]

Worse, herding will happen without any change in scientists' perceptions of the competing programs' returns functions. Evidently, such herding might strongly favor a program whose predictive merits are quite modest. What matters is that the contribution it makes is a predictive contribution; the quality of that contribution is secondary. Of course, if the program consistently fails to deliver on its promise, estimates of its potential will fall; this may in time undo the harmful effects of herding, as the attraction of the relatively unexplored and now relatively promising explanatory programs grows large enough to overwhelm worries about the recognition of explanatory achievement. But that process will take time.

Can this schema account for herding in economics? That question I leave to the historians. But the danger seems to me to be real.

How might it be averted? Some benevolent funding body might set up incentives providing additional motivation for scientists to join small, explanatory research programs—an "institute for new economic thinking"? Still there will be, on the other side, formidable discouragement: the worry that, whatever the inducements, even profound contributions to such programs will not earn economists' own applause—"the only coin worth having".

5. Even if the braking effect damps the bias, herding happens to a certain degree, since the braking cannot undo the fact that a greater proportion of scientists join the predictive program than is warranted by its intrinsic merits.

Acknowledgments

My thanks to the participants at the Workshop on Methodology, Systemic Risk, and the Economics Profession for many helpful comments and conversations over the course of that most fruitful weekend in December 2012.

References

Bikhchandani, S., D. Hirshleifer, and I. Welch. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy* 100:992–1026.

Dasgupta, P. and P. A. David. (1994). Toward a new economics of science. *Research Policy* 23:487–521.

Hands, D. W. (1997). Caveat emptor: Economics and contemporary philosophy of science. *Philosophy of Science* 64:s107–s117.

Kitcher, P. (1990). The division of cognitive labor. *Journal of Philosophy* 87:5–21.

Latour, B. and S. Woolgar. (1986). *Laboratory Life: The Construction of Scientific Facts*. Second edition. Princeton University Press, Princeton, NJ.

Merton, R. K. (1957). Priorities in scientific discovery. *American Sociological Review* 22:635–659.

Peirce, C. S. (1879). Note on the theory of the economy of research. In *Report of the Superintendent of the United States Coast Survey Showing the Progress of the Work for the Fiscal Year Ending with June 1876*, pp. 197–201. US Government Printing Office, Washington DC.

Samuelson, P. A. (1962). Economists and the history of ideas. *American Economic Review* 52:1–18.

Strevens, M. (2003). The role of the priority rule in science. *Journal of Philosophy* 100:55–79.