Comments on "Social Epistemology and Individual Rationality"
Lara Buchak, 5.21.2007
*note: these were written to be read aloud, and thus written somewhat colloquially

First of all, thanks so much to Ryan Muldoon for the opportunity to read about an interesting topic. I enjoyed the paper very much.

1. Summary of problem and Muldoon's simulations

The current rewards system in place for scientific advancement – that the first person to make a discovery gets all the credit and prestige, even if she is first by a matter of hours – seems to need an explanation. After all, there is a good deal of luck involved in who makes a discovery first, and wouldn't it make more sense to reward scientists based on effort, or some other scheme?

Kitcher and Strevens, in separate papers, provide an explanation for this rewards system: they claim that, if scientists are game-theoretically rational and care about rewards, then under this rule (which they call the "Priority rule"), they will distribute themselves optimally among research programs competing to solve the same problem. That is, rational scientists will distribute themselves as a central planner would distribute them among these research programs.

Muldoon focuses on Strevens' model, so I will briefly explain it.

Strevens allows that each research program has a probability of success that is a function of worker-hours put into the program; this function decreases marginally. Each program also has a return value; these are the same if the programs are trying to solve the same problem, namely the value of solving the problem. We are trying to maximize the *expected* return; that is, in the case we care about, we are trying to maximize the probability of *some* program solving the problem.

If the values of the programs were independent – that is, the success of one won't make the success of the other less valuable – then Strevens shows that the optimal way to distribute scientists is to put each scientist where she makes the greatest marginal contribution. The rule that leads game-theoretically rational scientists to distribute themselves this way is, simply, award prestige in proportion to the marginal contribution each scientist makes; then, as long as they know the probability of success functions and what other scientists are working on, they will position themselves where they will make the greatest marginal contribution. If the programs are *competing* – that is, if one succeeds then the other is of no value; this is the case we care about – then, Strevens shows, the rule that will lead scientists to distribute themselves optimally turns out to be the Priority rule!

This model, of course, relies on many assumptions, and Muldoon questions several of these. I'll talk about the two most crucial criticisms in my comments, both epistemological in nature. First, he criticizes the assumption that scientists know the intrinsic (which Muldoon takes to mean *objective*) probability functions. This seems to make them much more knowledgeable than is warranted. Second, he criticizes the assumption that agents know – and more specifically, have common knowledge of – how other scientists are distributed among projects. He shows that if we relax this second assumption, scientists will distribute themselves non-optimally. To show this, he runs simulations in which each agent only knows which projects his 'neighbors' are working on, and picks a project that will optimize his expected returns on the basis of only this information. On Muldoon's simulations, if the neighborhood is too small, then under the Priority Rule (and the marginal contribution rule), scientists do not distribute themselves optimally; in fact, they all tend towards the project with the highest antecedent probability of success. And Muldoon shows that the neighborhoods need to be quite large.

He proposes modeling the problem of how scientists decide which projects to work on as a hill-climbing problem instead.

2. Defending Strevens against the criticisms

I find no fault with Muldoon's simulation data; it is interesting to see how an optimal distribution of scientists depends on how much they know about what each other is doing.

However, I can offer a defense of Strevens against Muldoon's two criticisms:

But first, a brief response to Muldoon's criticisms of the non-epistemic assumptions. As for the claim that the assumption of risk-averse scientists undermines the project: you could think that scientists, when choosing a *rule*, are motivated by scientific advancement – or that the scientists have no choice about what the rule is, because that is up to a larger section of society. And then self-interest (and risk aversion about a scientist's own fortune) comes into play in project selection. As for the question about the ratio between scientists and projects: who becomes a scientist is probably driven by the market more generally, so it's a larger question; but for Strevens' project, we just care about the optimal distribution of the labor we have, anyway.

(1) First, Muldoon's criticism that scientists know the "intrinsic" probability functions:

Contra Muldoon, we're not really talking about objective probabilities (chances). What we care about is "how do scientists distribute themselves when certain 'payoff structures' are in place?" which depends on the *subjective probabilities (or probability functions)* scientists assign to the success of various projects, and we care about "how would a central planner distribute scientists?" which depends on her subjective probability function (presumably the subjective probability function of a representative individual or the like).

In other words, what Strevens claims is that, under the priority rule, scientists distribute themselves optimally, *given what everyone believes* will solve the problem. He does not claim that they distribute themselves optimally given which methods will *actually* solve scientific problems.

Now, it does seem, on the face of it, that we need the subjective probability functions of the scientists to agree, since we're using a "representative agent." And maybe this is too strong of an assumption. But can we relax this assumption, and still get the result we want, namely that under the priority rule, scientists would distribute themselves as a central planner would?

For example, let's assume that the central planner's subjective probability function is the average of the scientists' subjective probability functions. (This preserves talk about expected success.) Then, as long as the scientists with probability functions very far from the average (say, the scientist who assigns a high credence to the success of receiving a cure for cancer in his dreams) distribute themselves first, scientists will still distribute themselves roughly optimally. And it would make sense that probability distributions of later scientists would be closer to the average, since more knowledge or evidence should lead to convergence.

Or so the intuition goes. But Muldoon's data on different subjective beliefs does not bear this out. Why? One answer to this question highlights an interesting and perhaps controversial feature of Muldoon's model. On his model, scientists don't just show up, pick a project, and stay there. Rather, each 'round,' they evaluate what others are doing and respond. So the turnover rate for projects can be extremely high. Thus, scientists are all doing their part to force the distribution that is optimal from their point of view, including the scientists with outlying beliefs. Whether the high turnover is warranted may depend on empirical facts about the practice of science, and it would be interesting to know these facts.

(2) Now to Muldoon's criticism of Strevens' assumption that scientists know how other scientists distribute themselves among projects. I agree that Muldoon shows that an assumption *like* this is important, but we need to figure out what that assumption is.

First of all, agents do not need any kind of knowledge of each other's *beliefs*, unless they are all deciding which project to work on before seeing what anyone else is working on; but this is clearly not intended in the model, and Muldoon's use of dynamic modeling also belies this assumption. So they certainly don't need *common* knowledge of each other's beliefs. They don't need to know how other scientists will distribute themselves among projects ahead of time; they just need to know how scientists are distributed at the time they themselves sign on to a project.

But, as Muldoon points out, maybe the assumption that scientists know how others are distributed is too strong; is it plausible to think that scientists know exactly how many other scientists are working on each project?

As it turns out, all Strevens requires is a much weaker assumption: that scientists *behave as if* they know how other scientists are distributed among projects. This might be a plausible assumption, although it takes some work to defend. But the intuition behind it is that even though scientists may never go out and look at the exact numbers, they may have a general sense of when a project is being over-worked already and when they're working on something few people have considered. Or they might have an intuitive grasp of how much of a contribution they can make to various projects without knowing how many people are working on them because they know what's already been done and what hasn't been done on a particular project. In fact, all scientists really need to know (or estimate) is their potential marginal contribution to each project, and the (subjective) probability of the success of each project; knowing what other agents are doing is just one way of knowing this.

Still, let's grant Muldoon's point: in order for the priority rule to cause rational scientists to distribute themselves *optimally*, they must have a good idea of the distribution of scientists, or, in Muldoon's terms, the radius of vision must be large enough.

But even if Muldoon can show empirically that scientists usually don't know what other scientists are working on (that the radius of vision is small), this doesn't refute Strevens claim that the priority rule is *better than other rules* (and hence, that the priority rule can be explained by reference to game-theoretic rationality). It could just be that it's very hard to achieve an optimal distribution when no one knows anything about anyone else's work (this wouldn't be surprising), but that the priority rule helps a little bit. Or it could be that the priority rule is on a par with other rules in these situations, but sometimes scientists do have more information, and it's good to have the priority rule in place for those times.

One final note before I move on: I have been defending Strevens in some places by pointing out that all he needs is *approximate* rationality. It isn't always true that a result for rational people will be approximately true for approximately rational people; a small deviation in rationality could lead to a huge difference in results. However, in this case, we're dealing with smooth, continuous functions, so approximately rational scientists will approximate the optimal distribution (as long as we don't deviate too far, for example make the radius of vision small).

Of course, Muldoon's point about a small radius of vision could show that we need a different model. I'll briefly discuss Muldoon's suggestion that we use a hill-climbing model instead.

3. Comments on Muldoon's suggestion of a hill-climbing model

The original question that Strevens set out to answer was: why does the priority rule make sense? And the answer he came up with was: under the priority rule, rational self-interested scientists will distribute themselves so that we have a higher expectation (given the beliefs we have) of solving problems than we will under other rules.

Muldoon claims that this is only true if we make some assumptions, notably that scientists (roughly) agree on the probabilities of success, and that scientists know what many other scientists are working on. I think he's done a defensible job of showing this latter assumption is important. And he's proposed an alternate way to think about the way scientists distribute themselves: as a hill-climbing problem, possibly assigning different landscapes to different scientists, depending on what they value.

Back to the original question. How do we represent the priority rule on the hill-climbing model? Here, the landscape itself can represent the reward scheme: as Muldoon told me when we were corresponding about this, as a scientist settles on a patch of land, she takes the reward. In the priority scheme, the first person to settle on a peak takes all of the value, whereas in, say, a labor-based scheme, value depends on how long a scientist settles at a peak. It is still an open question whether or not the priority scheme will maximize the chance of solving certain difficult problems – or what sort of distribution of scientists various reward schemes will lead to. This seems like a hard problem – since we're adding so many parameters – but also an exciting opportunity for further research, and I look forward to seeing some of this research from Muldoon in the future.

Of course, this has now changed the question substantially. If we're scrapping the original question because we're scrapping Strevens' model, how does Muldoon explain the apparent success of Strevens' explanation for the priority rule? As I've said, maybe the new model can explain it, but this remains to be seen.

I will close by saying that I look forward to more research in this area.