ARTICLE IN PRESS

Studies in History and Philosophy of Science xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Studies in History and Philosophy of Science

journal homepage: www.elsevier.com/locate/shpsa

The credit incentive to be a maverick

Remco Heesen[a,b,*]

[a] Department of Philosophy, School of Humanities, University of Western Australia, Crawley, WA, 6009, Australia
[b] Faculty of Philosophy, University of Cambridge, Sidgwick Avenue, Cambridge, CB3 9DA, UK

HIGHLIGHTS

- Defines mavericks as impact-seekers, scientists going in for high-risk high-reward research, in contrast to safety-seekers.
- Provides a model of a three-way tradeoff between impact, success probability, and speed.
- Rational credit-maximizing scientists do not form particularly effective communities in this model.
- An escape route from this result and its limitations are discussed.

ABSTRACT

There is a commonly made distinction between two types of scientists: risk-taking, trailblazing mavericks and detail-oriented followers. A number of recent papers have discussed the question what a desirable mixture of mavericks and followers looks like. Answering this question is most useful if a scientific community can be steered toward such a desirable mixture. One attractive route is through credit incentives: manipulating rewards so that reward-seeking scientists are likely to form the desired mixture of their own accord. Here I argue that (even in theory) this idea is less straightforward than it may seem. Interpreting mavericks as scientists who prioritize rewards over speed and risk, I show in a deliberatively simple model that there is a fixed mixture which is not particularly likely to be desirable and which credit incentives cannot alter. I consider a way around this result, but this has some major drawbacks. I conclude that credit incentives are not as promising a way to create a desirable mixture of mavericks and followers as one might have thought.

1. Introduction

It is an old idea in the philosophy of science that there are different ways to contribute to science. Kuhn famously distinguished between normal science and extraordinary or revolutionary science, and he strongly hinted that scientists were differentially suitable to these different types of work (Kuhn, 1996, pp. 82–89). Hull invokes a similar idea, and makes the tie to individual scientists more explicit: "Most scientists play it safe, making small, uncontroversial contributions. A few attempt to revolutionize some area of science. They risk failing big" (Hull, 1988, p. 474).

Let me unpack this distinction from Hull a bit. Many scientists work in the context of established research programs or paradigms, using standard methodologies to make incremental progress on relatively specific pre-existing problems. Generally, they are able to produce results that are likely to be accepted by the scientific community, and do so consistently over time. Some scientists instead go in for research in relatively unexplored areas, or areas where the foundations are not as settled. This type of work can yield major new discoveries if successful, but it is more likely to take an unpredictable amount of time or fail completely. For the moment I will refer to scientists of the latter type as *mavericks* and to the former as *followers*.

The distinction has drawn some recent interest in the literature on *formal social epistemology*, i.e., the study of epistemological questions regarding the social structure of science using formal models. Weisberg and Muldoon (2009) introduce *epistemic landscape* models and use them to argue that a scientific community benefits from having a mixture of mavericks and followers. A number of subsequent papers have extended or criticized this (Grim et al., 2013; Alexander, Himmelreich, & Thompson, 2015; Thoma, 2015; Pöyhönen, 2017; Avin, forthcoming).

One of the key underlying questions in these papers is: what *distribution of researcher types* (i.e., what proportion of mavericks, followers, and possibly other types) makes for an effective scientific community? Here, an effective community appears to be one that finds and uses approaches that work well for the research question(s) it is interested in.

The authors involved do not take this question about good (or optimal) distributions to be of purely theoretical interest. Rather, once a satisfying answer to the question is obtained, it should be used normatively: real scientific communities should be encouraged to adopt a

* Corresponding author. University of Western Australia, Australia.
E-mail address: remco.heesen@uwa.edu.au.

https://doi.org/10.1016/j.shpsa.2018.11.007
Received 13 December 2017; Received in revised form 12 June 2018; Accepted 28 November 2018
0039-3681/ © 2018 The Author. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/BY/4.0/).

Please cite this article as: Heesen, R., Studies in History and Philosophy of Science, https://doi.org/10.1016/j.shpsa.2018.11.007

ARTICLE IN PRESS

good (preferably optimal) distribution of researcher types. In the words of Kitcher, one of the first to write on formal social epistemology: "The philosophers have ignored the social structure of science. The point, however, is to change it" (Kitcher, 1990, p. 22).

On the whole, the papers mentioned above say very little about how a socially beneficial distribution of researcher types, once determined, should be realized. Thoma (2015) worries about this a little bit. Towards the end of her paper, she points out that in Weisberg and Muldoon's model, it is "unclear why anybody would choose to be a follower, given their lack of productivity" (Thoma, 2015, p. 470). In her own model, where explorers and extractors play the role of mavericks and followers respectively, a converse problem arises: "the question is why anybody would choose to be an explorer" (p. 470). She then suggests that rewards given to individual scientists may have a role to play here: "social and financial rewards…could help to maintain an epistemically beneficial diversity of research strategy by making sure that the explorer strategy is attractive enough for some scientists to choose it" (p. 471).

This idea has some plausibility. It is a well-known fact that scientists are mainly rewarded for their work through *credit*: the recognition of achievements by peers, made visible in the form of publications in prestigious journals, citations, prizes, and prestigious appointments (Hull, 1988; Merton, 1957, 1973). This affects scientists' behavior in a predictable way: they are more likely to do things which they believe will yield credit (Dasgupta & David, 1994; Merton, 1969; Strevens, 2003). Formal social epistemologists like Kitcher (1990) and Strevens (2003) have suggested that this may affect scientists' choice of methodological approach—creating a socially beneficial *division of cognitive labor* in Kitcher's terminology. A number of authors have argued for other positive side effects of the credit motive (Boyer-Kassem & Imbert, 2015; Dasgupta & David, 1994; Heesen, 2017; Zollman, 2018).

The question, then, that I focus on here is whether *credit incentives* can be used to encourage scientists to create a socially beneficial distribution of researcher types. In other words, where previous work has asked what a good distribution of researcher types looks like, I ask whether credit can be used to get there. I use a deliberately simple model (abstracting away from epistemic landscapes) as a starting point for my discussion.

I claim that under quite general conditions the key factor that influences whether a credit-maximizing scientist chooses to be a maverick or a follower is her *predisposition* for producing the type of work typical of a maverick or a follower. I use this to argue for a negative conclusion: there is no reason to think that the predisposition to be a maverick occurs among scientists at anything like the frequency at which it needs to occur to produce a particularly good distribution of researcher types.

As will become clear, I do not intend to make any strong nativist claims about predispositions. In particular, I remain agnostic about when and how a scientist's predisposition gets fixed (at birth, in childhood, or as a part of professional training). However, it is key to my argument that this predisposition is specific to an individual scientist and is unlikely to change much over the course of a scientific career.

My argument uncovers serious limitations to the idea of using credit incentives to create a beneficial distribution of researcher types. Credit incentives do not act like an "invisible hand", steering the scientific community toward the optimal distribution, as they have been claimed to do for the division of cognitive labor (Strevens, 2003) and intermediate results sharing (Heesen, 2017).

Not all is lost though. After presenting my formal result, I discuss in more detail what kinds of manipulation through credit incentives are ruled out by it, and what remains possible. The relationship between the division of cognitive labor (studied by Kitcher and Strevens) and the distribution of researcher types (studied by Weisberg and Muldoon and Thoma) turns out to be more complicated than previously thought. Where Weisberg and Muldoon and Thoma saw an analogy, I argue there is a tradeoff instead.

2. Mavericks as impact-seekers

What exactly is a maverick? The various authors mentioned in the introduction all use slightly different definitions so it is useful to clarify the version of the idea I work with.

Weisberg and Muldoon put their definitions in terms of a desire to deviate from (or imitate) what others have done: "mavericks avoid previously examined approaches, while followers emulate them" (Weisberg and Muldoon, 2009, p. 243). In a variation on this theme, Thoma's explorer types "like to follow approaches that are very different from those of others, while extractor-types like to do work that is very similar to but not the same as that done by others" (Thoma, 2015, p. 463). No analysis is given of what these desires are based on.

Kuhn (1996) does not explicitly discuss different types of scientists, but associates different skills with periods of normal science as opposed to extraordinary or revolutionary science. His most explicit claim in this regard, repeated multiple times (pp. 90, 144, 166), is that younger scientists are more suitable to instigate revolutions (cf. Wray, 2003).

Hull (1988, p. 474) draws the distinction in terms of the tradeoff between risk and reward. This way of drawing the line is most congenial to my purposes. It avoids building into the definition any assumption about a psychological explanation for the existence of different researcher types, and it also avoids building in an assumption that the distinction has to be binary (as opposed to graded).

So on my definition, a "maverick" is a scientist (used broadly to include any academic researcher) who works on high-risk high-reward projects, i.e., a scientist who aims "to revolutionize some area of science" (Hull, 1988, p. 474). This presupposes that scientists are in a position to decide in advance whether to attempt a high-risk high-reward project as opposed to something less ambitious. An opposing view holds that scientists pick projects for other reasons (e.g., intellectual interest, or the availability of instruments or materials) and that high-reward projects are simply regular projects carried out by particularly skillful or lucky scientists. I do not provide an argument against this view, but I assume that, at least, the truth is somewhere in the middle: scientists can form expectations about the likely rewards of particular projects and such expectations factor into their decisions. Let me say more now about what I mean by rewards and risks.

As mentioned, the main unit of reward in science is *credit*: prestige conferred by the scientist's peers in the form of prizes, publications, citations, and appointments. Founding or revolutionizing a (sub)field of science is the highest possible achievement, and as such yields the most credit. Contributions that change a (sub)field in some smaller way yield less credit. Whereas scientific work that is completely ignored sits at the bottom and yields no credit. Hence credit is given for a contribution to science broadly in proportion to its scientific *impact*.

That credit is proportional to impact is an empirical claim which I build into my model. It is defended by Merton (1957, especially pp. 642–645) who argues that "rewards are to be meted out in accord with the measure of accomplishment" (p. 659). Cole and Cole (1967, p. 385) provide some limited quantitative evidence by showing that scientists who write high-impact papers tend to get prestigious awards and have better name recognition. It is perhaps telling that both credit and impact tend to be measured using citation counts (e.g., Sinatra, Wang, Deville, Song, & Barabási, 2016, Cole & Cole, 1973, with the latter using citations in both senses; see Bornmann & Daniel, 2008 and citations therein for discussion). I discuss what happens if this assumption is relaxed in section 4.

The risk of aiming for high impact is twofold. First, the project may fail, as most attempts to revolutionize an area of science do. (In contrast, less world-shocking projects often have a greater chance of success.) Second, projects that aim for high impact tend to take more time to complete. In some sense this exacerbates the first risk: if the project fails, potentially a lot of time has been wasted, at least compared to the time wasted when a less ambitious project fails. But even if the project succeeds the risk of taking a long time is relevant: if the scientist could

ARTICLE IN PRESS

have completed multiple less ambitious projects in the same amount of time, the combined credit from those projects might have been higher than that of the high impact project, even if it is more impactful than any one of the less ambitious projects.

So while Hull describes a two-way tradeoff between risk and reward, I suggest that the risk component of this tradeoff may be broken down into a risk of failure and a risk of being slow (which carries an opportunity cost). The next section discusses a simple model of the three-way tradeoff between what I will call impact, speed, and success probability. Insofar as I continue to use the terminology of risk, it will refer to the risk of failure (the complement of the success probability) rather than the risk of being slow.

To distinguish my conception of different researcher types from that of authors like Weisberg and Muldoon and Thoma, I will use different terms from here on out. Emphasizing Hull's tradeoff, I refer to a scientist who works on high-risk high-reward projects as an *impact-seeker* (my analogue of the maverick) and to the contrast case as a *safety-seeker*.

3. A simple model of researcher types

The key tradeoff that the model aims to capture is the three-way tradeoff between impact, success probability, and speed. The model I provide is deliberately simple, as the aim is to capture the phenomenon I am interested in with as few assumptions as possible.

As the old business adage says: you can have it good, fast, or cheap; pick two. In the present context, this means you can have it impactful, fast, or with high chance of success, and (within certain limits) you get to choose two of these, with the third one being determined as a result of the tradeoff. In setting up the model, I assume that success probability and impact are chosen by the scientist, and speed is determined as a function of these choices. (This is a harmless assumption: a simple translation exercise yields the same model with success probability and speed as the independent variables, or impact and speed.)

Let p denote the scientist's *success probability*. More precisely, this is the scientist's subjective probability that the project yields a publishable result of any impact at all. Being a probability, this variable is constrained to the unit interval: $p \in [0,1]$. The risk of failure is the complement $1 - p$ of the success probability. Note that what matters to the scientist's choice of what to work on is her own estimate of the likelihood of success, hence this is a credence rather than a chance. It seems reasonable to suppose that a competent scientist's credence of success cannot be too far off from her (objective) chance of success, but nothing in my model turns on this.

Let c denote the scientist's *impact*. This is the scientist's estimate of the impact of her project assuming it succeeds. As discussed in the previous section, in the model I equate impact and credit. I remain agnostic about how impact/credit should be measured, but one relatively simple way to make this more precise would be as the expected number of citations to the publication(s) resulting from the project, as is commonly done when a quantitative measure is needed (Cole & Cole, 1967, 1973; Sinatra et al., 2016). I assume estimated impact to be nonnegative (otherwise why start the project?), i.e., $c \in [0, \infty)$. I do not explicitly impose a maximum on the impact a scientist might aim for, although, as I will show, the assumptions below entail that such a maximum exists.

Finally, let λ denote the scientist's *speed*. The scientist has to estimate how long her project will take with a probability distribution. Empirical data on publication patterns in scientific careers suggests that the length of projects approximates an exponential distribution (Huber, 1998a, b, 2001; Huber & Wagner-Döbler, 2001a, b). So I assume the scientist believes the duration of the project to be exponentially distributed, with the speed parameter λ denoting her "work rate". That is, λ gives the average number of projects she would complete per unit of time if she works on projects at rate λ continuously. A more straightforward interpretation attaches to its inverse: $1/\lambda$ is the expected completion time of the project. As with the other two variables, I focus on the scientist's own subjective estimate of the speed.

The tradeoff is modeled by assuming that speed is determined as a function of success probability and impact: for any p and c, $\lambda(p, c)$ denotes the maximum speed the scientist thinks she can achieve given her choices of success probability and impact. Call this function *the tradeoff function*. I make the following assumptions on this function.

Assumption 1. The tradeoff function has the following properties:

1.a. The tradeoff function is decreasing in both of its arguments, i.e., if $p < p'$ then $\lambda(p', c) < \lambda(p, c)$ for any c, and if $c < c'$ then $\lambda(p, c') < \lambda(p, c)$ for any p.

1.b. No perfect work: $\lim_{p \to 1} \lambda(p, 0) = 0$.

1.c. The tradeoff function is concave: for any $t \in [0,1]$ and for any p, p', c, c'

$$t\lambda(p, c) + (1 - t)\lambda(p', c') \leq \lambda(tp + (1 - t)p', tc + (1 - t)c').$$

Assumption 1.a formalizes the fact that the tradeoff function is meant to capture a tradeoff: if one of the three variables goes up, one or both of the others must go down. This need not always happen in practice, but recall that the variables reflect the scientist's expectations before starting the project. The assumption only fails if there is a way that the scientist can know about in advance for her to increase one of these variables without a corresponding decrease. This seems like it would be a rare case.

Assumption 1.b captures the idea that there is no certainty in science. There is never a guarantee of success, hence impact and speed must go to zero as the success probability goes to one.

Assumption 1.c indicates that there are decreasing marginal returns from decreasing a variable or a combination of two variables, or increasing marginal costs for increasing a variable. A scientist aiming to increase her success probability requires ever more time to do so (or must restrict her ambitions of impact). A scientist who wants to publish more quickly finds that her gains are smaller if she already had relatively modest goals for impact and success probability (if only because writing a paper itself takes time, and this will take up a relatively larger share of the time if the scientist spends less time on developing the scientific content). And a scientist going for greater impact faces ever greater increases in the risk of failure or the required time.

Note that these assumptions do not specify the exact rate at which the three variables trade off. In particular, there is no assumption of symmetry. Thus, it is consistent with my assumptions that aiming for high impact comes at a cost of low chance of success *and* low speed, whereas perhaps high chance of success and high speed can be achieved simultaneously (given a specification of what counts as "high" for each variable). The substance of it being a three-way tradeoff is in the fact that a small increase in impact (say) can usually be "paid for" either with a decrease in success probability, a decrease in speed, or a combination of both.

These assumptions have a number of consequences. One is that the tradeoff function is continuous, i.e., there are no sudden jumps (this follows from concavity; see Lemma 1 in Appendix A). This seems right to me for most cases, as the tradeoff function reflects the scientist's expectations about average speed, not the actual time taken on any particular project. Even so, I grant that there may be exceptions to this; my aim is to capture typical cases, not necessarily all.

Another consequence pertains to the set of choices for p and c for which the tradeoff function is nonnegative. This may be viewed as the set of feasible choices. Assumption 1 entails that there is a function μ which may be called the *maximum impact function*: for any choice of p, $\mu(p)$ gives the highest possible impact that can be achieved at success probability p and nonnegative speed (see Appendix A for a proof).

Proposition 1. (*Maximum Impact*). *Let* $p \in [0,1]$. *Assumption* 1 *entails that there exists a function* μ: $[0,1] \to [0, \infty)$ *such that* $\lambda(p, c) \geq 0$ *if and only if* $c \leq \mu(p)$. *The function* μ *is decreasing and concave.*

ARTICLE IN PRESS

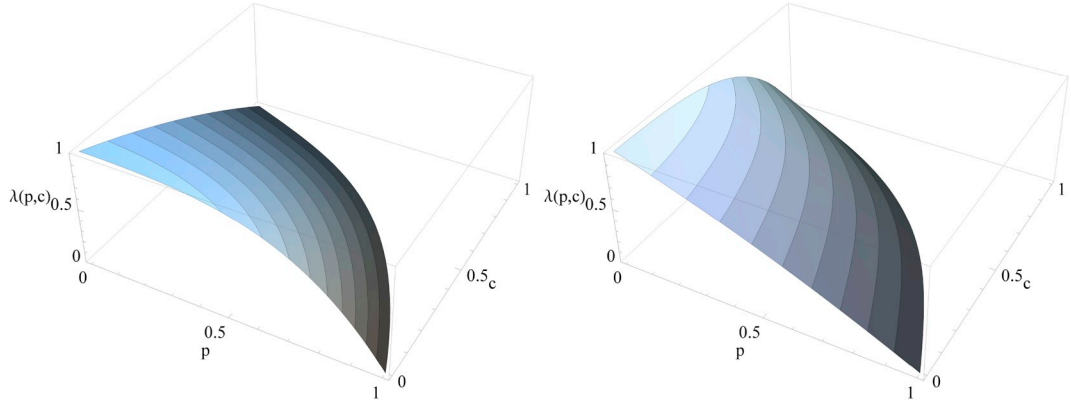

Fig. 1. Graphs of λ_1 (on the left) and λ_2 (on the right).

Note that the maximum impact function inherits the key properties of the tradeoff function. This reflects the more general fact that, if a particular level of speed is fixed, the two-way tradeoff between impact and success probability inherits (through Assumption 1) the structural properties of the three-way tradeoff.

Consider a scientist who aims to *maximize expected credit*, and is perfectly rational in going about this goal. This is not to claim that any such scientist exists: real scientists have other goals beside credit, and are not perfectly rational. But the choices a rational expected credit-maximizing scientist makes are the choices that real scientists have a credit incentive to make.

The scientist's expected credit per unit of time while she is working on this project is a function of her choices of success probability p and impact c. She completes projects like this one at a rate of λ per unit of time, the project succeeds with probability p, and the scientist's expected credit if the project does succeed is c. Hence the expected credit function C is given by

$$C(p, c) = c{\cdot}p{\cdot}\lambda(p, c).$$

Note that this expected credit function does not depend on whether there are any other scientists in the community and what they are doing. It turns out to be a consequence of using exponential distributions to model the time it takes to complete projects that the *average* completion rate of projects is independent of the activity of competing scientists. This is mathematically convenient here as it reduces the model to one that can be analyzed using decision theory rather than requiring game theory. It is not obvious that anything significant is lost by making use of this convenience, given the evidence that project completion times are indeed exponentially distributed (Huber, 1998a, b, 2001; Huber & Wagner-Döbler, 2001a, b).

So what happens when the scientist aims to maximize the expected credit function?

Theorem 1. (*Unique Maximum*). *If* Assumption 1 *is satisfied, there exists a unique point* (p^*, c^*) *that maximizes the function C:*

$$C(p^*, c^*) = \max_{p\in[0,1], c\in[0,\infty)} C(p, c).$$

Moreover, $0 < p^* < 1$ *and* $0 < c^* < \mu(p^*)$.

The key insight offered by Theorem 1 (proved in Appendix A) is that there is a *unique* credit-maximizing choice. In other words, Theorem 1 rules out the possibility that a scientist could switch from being a safety-seeker to an impact-seeker or vice versa, while remaining at a global maximum of C. For a credit-maximizing scientist, there is just one rational tradeoff between impact and safety, not a range of admissible values between which an independent preference for being an impact-seeker or a safety-seeker might act as a tie-breaker. This consequence of the model may be seen as surprising in light of the psychological terms in which previous authors have discussed these types (this is most explicit in Thoma, 2015, for example on p. 470: "some scientists do choose to be explorers, apparently for good reasons").

This does not rule out the existence of different researcher types. But it suggests that these types are the result of differences in the shape of the tradeoff function of different scientists. According to Theorem 1, the location of the optimum given a particular tradeoff function determines the researcher type any scientist with that tradeoff function has a credit incentive to be.

There is some reason to believe that the tradeoff function is more or less fixed over the course of a career. For example, Huber (1998a, b, 2001) and Huber and Wagner-Döbler (2001a, b) argue based on empirical evidence that the productivity of scientists (i.e., λ) is, on average, constant over the course of a career. And Sinatra et al. (2016) argue that high-impact papers are distributed randomly within each scientist's career, suggesting that expected impact (i.e., c) is constant across a career.

This suggests an interpretation of the tradeoff function as a relatively stable *predisposition* to trade off speed, success probability, and impact in a particular way. Any given scientist's researcher type will then similarly be relatively stable, as determined by her predisposition, while different researcher types result from differences in predispositions across scientists. The following Example illustrates this.

Example. Consider two scientists. In the estimation of scientist 1, the tradeoff between speed, success probability, and impact is given for her by the function λ_1, where

$$\lambda_1(p, c) = 1 - \frac{3}{4}p^4 - \frac{1}{4}p^2 - \frac{1}{2}pc - \frac{1}{4}c^2 - \frac{3}{4}c,$$

for all $0 \le p \le 1$ and $c \ge 0$ (see Fig. 1). Note that this function satisfies Assumption 1. Then the credit-maximizing choice for scientist 1 is $p \approx 0.52$ and $c \approx 0.38$.

In contrast, scientist 2's estimate of the tradeoff is given by tradeoff function

$$\lambda_2(p, c) = 1 - \frac{1}{4}p^2 - \frac{1}{2}pc - \frac{1}{4}c^2 - \frac{3}{4}c^4 - \frac{3}{4}p,$$

for all $0 \le p \le 1$ and $c \ge 0$ (see Fig. 1). This function also satisfies Assumption 1. But the credit-maximizing choice for scientist 2 is $p \approx 0.38$ and $c \approx 0.52$.

Scientist 1's tradeoff function shows a predisposition that favors the success probability compared to scientist 2's, which shows a predisposition that favors impact. This is because λ_1 is closer to linear in c—having only a small quadratic component—while it is a fourth-degree polynomial in p (λ_2 is simply its mirror image). So if the scientists are responsive to credit incentives, scientist 1 will behave more like a safety-seeker, doing relatively safe, low-impact research. Scientist 2 on the other hand will behave more like an impact-seeker, doing more risky, high-impact research.

If C correctly captures the shape of the credit function, and

ARTICLE IN PRESS

Assumption 1 is justified, then Theorem 1 guarantees that differences in the shape of the tradeoff function (what I am now calling different predispositions) are the *only* way different researcher types can arise as a result of credit incentives.

It is tempting to think of a scientist with a predisposition that favors safety-seeking as risk-averse, and a scientist with a predisposition that favors impact-seeking as risk-loving, since the latter will more likely engage in high-risk high-reward research. But this is somewhat misleading. At least on a conventional understanding of the terminology, risk aversion is thought of as a preference. A preference to avoid risk can be expressed by using a concave utility function (setting aside complications introduced by the Allais and Ellsberg paradoxes, although these types of risk aversion are also most readily understood as preferences).

In contrast, the predispositions discussed here should be thought of as beliefs. The tradeoff function captures a scientist's belief about how speed, success probability, and impact trade off for her. So a predisposition that favors safety-seeking reflects a scientist's belief that her knowledge and skills will generate more impact over time if she focuses on a larger number of low impact projects than if she focuses on a smaller number of high impact projects.

I have said nothing about whether the distribution of predispositions in a scientific community is likely to incentivize a good or a bad distribution of impact-seekers and safety-seekers. As far as I can see, there is no connection between how predispositions will be distributed and what makes for a good distribution of researcher types, so any positive effects here would be purely coincidental. Moreover, according to Theorem 1, credit incentives cannot be used to *improve* the distribution of types. Rather, the distribution of types that is favored by credit incentives—and hence the distribution the community will tend to insofar as scientists are responsive to credit incentives—is as likely to be beneficial as harmful.

4. Possibilities and limitations

As mentioned, I take the tradeoff between speed, success probability, and impact as it is captured in the tradeoff function to be a feature of an individual scientist. I find the term *predisposition* convenient as a shorthand, but the implications of this terminology should not be taken too strongly. As mentioned above, the idea is to capture something that is specific to an individual scientist and unlikely to change (much) over time.

Theorem 1 tells against a “passive” approach to using the credit economy to get a good distribution of researcher types. Where previous work has argued that the credit economy operates like an invisible hand, using individual self-interest to motivate socially beneficial decisions (Heesen, 2017; Hull, 1988, 1997; Leonard, 2002; Strevens, 2003), I have argued here that this mechanism should not be expected to have particularly beneficial effects in the context of the distribution of researcher types. This is because there is no reason to expect the distribution of predispositions in a scientific community to match the optimal (or a particularly good) distribution of impact-seeking versus safety-seeking.

This does not rule out a more “active” approach. The way scientific results are rewarded is to some extent under our control. More specifically, it is under the control of “gatekeepers” such as journal editors and the agencies and governments that award grants and prizes. Scientists’ credit incentives change if the amount of credit that is given for different scientific contributions changes.

So far I have assumed that scientists expect the reward for successful scientific work to be directly proportional to its scientific importance: both are represented by the variable c. Scientific importance (i.e., impact) influences the tradeoff with speed and success probability and thus features in the tradeoff function, whereas the reward features in the credit function. That rewards are proportional to importance might be a feature of the credit economy as currently implemented (see Cole & Cole, 1967; Merton, 1957; Strevens, 2003), but if it is, it is not a necessary feature. Proportionality could be abandoned.

Suppose a social planner (with perfect control over how rewards are distributed) observes that not enough scientists are working to revolutionize the field with high-risk high-reward research, i.e., there is a relative shortage of impact-seekers. Such a planner could start giving extra rewards for high-impact work, perhaps by introducing new prizes for this kind of work. As a result, the expected credit per unit of time for individual scientists looks different. For example, the result might be that credit per unit of time is described by a new function C' given by

$$C'(p, c) = c^2 \cdot p \cdot \lambda(p, c),$$

i.e., the scientist now expects rewards to be proportional to the square of the scientific importance of the work, meaning that highly important work now gets rewarded more (relative to the original credit function C). In general, this will shift the optimal tradeoff for her so she becomes more impact-seeking. To illustrate this: the C'-maximizing choice for scientist 1 in the Example above has $c \approx 0.53 > 0.38$ and the C'-maximizing choice for scientist 2 has $c \approx 0.64 > 0.52$.

This shows that, despite the result from the previous section, credit incentives can be used to manipulate the tradeoff scientists make between impact and safety. But then what, if anything, is the significance of the result from the previous section?

First, even allowing for the kinds of manipulations just suggested (which assume a kind of fine-grained control over rewards that may not exist in practice), credit incentives will not affect the *relative* distribution of impact-seekers and safety-seekers. Rewarding high-impact work more (or less), will generally affect all scientists in the same way. In the example above, all scientists become more impact-seeking, at least qualitatively speaking.

At the same time, the quantitative effects may be different. In the example, c increases by 0.15 for scientist 1 and 0.12 for scientist 2. Interestingly, it is the relatively more safety-seeking scientist 1 who makes a greater change in the direction of impact-seeking. So at least in this example, encouraging more impact-seeking also brings the scientists' choices of c closer together (here again there is a kind of decreasing marginal returns). While I suspect this particular point does not generalize, it illustrates that quite a detailed understanding of scientists’ predispositions may be needed on the part of a social planner to achieve the desired effect from these manipulations.

More specifically, if two scientists have the same predisposition, changing the rewards as suggested above will have the same qualitative *and* quantitative effect on both. So if the optimal distribution of researcher types requires changing the *diversity* in the scientific community (as Weisberg and Muldoon 2009 seem to suggest; but see Alexander et al., 2015; Thoma, 2015, and Pöyhönen, 2017 for critical discussion) it will not be possible to achieve this. For example, suppose all scientists in a given community have one of two tradeoff functions (say λ_1 and λ_2 from the Example). If the overall productivity of the community is optimized when there are three researcher types (e.g., impact-seekers, speed-seekers, and success-seekers) then credit incentives cannot produce the optimal distribution. Or to stay a bit closer to Weisberg and Muldoon's discussion, if the optimal distribution calls for 60% impact-seekers and 40% safety-seekers, but the community contains 30% impact-seekers (scientists with tradeoff function λ_2) and 70% safety-seekers (tradeoff function λ_1), the optimal distribution cannot be reached through the approach proposed here.

Second, the result of the previous section addresses the special case where expected credit is proportional to impact. Strevens (2003) argues that credit *should* be proportional to impact. This argument, if accepted, yields a preference for credit function C over C' or any alternative. Relatedly, Strevens (2006) argues that the Matthew effect—famous scientists receive more credit for the same work than non-famous scientists—is good for science on the assumption that credit is proportional to impact. While I will not consider this argument in detail, it supports the same preference. Any call to change the credit function

will need to take all factors into account.

The argument in Strevens (2003, section 2) is based on the following simple model. Suppose scientists need to choose which of a number of research problems to work on, and assume the following. First, the projects are independent in the sense that solving any one of them would be socially valuable regardless of which of the other ones are solved (Strevens calls this "the additive case"). Second, each project is characterized by a "success function", which specifies the chance of solving the problem given the number of scientists working on it. And third, the success functions are increasing and concave. Then a reward structure which gives credit to each scientist proportional to the expected impact of their contribution will incentivize scientists to distribute themselves among the problems in a way that maximizes the expected sum total impact.

The problem of how to distribute scientists over different research problems (or methodological approaches, or research programs) has been called the *division of cognitive labor* since Kitcher (1990). The question of the type of research carried out by each scientist (impact-seeking or safety-seeking) is a separate issue, one that I have here referred to as the *distribution of researcher types*; Thoma calls this the *diversity of research strategy*. Using this terminology, Strevens (2003) argues that giving credit proportional to impact optimizes the division of cognitive labor, and I have argued that there is no reason to expect it to lead to a particularly good distribution of researcher types. If the work of Weisberg and Muldoon (2009), Thoma (2015), and others yields a specific recommendation for a good (or even optimal) distribution of researcher types, incentivizing this distribution would presumably involve a departure from setting credit proportional to impact (i.e., replacing the function C, perhaps with the function C'). But this entails a departure from the incentive structure that optimizes the cognitive division of labor according to Strevens.

So there is not quite a disagreement with Strevens here. Strevens (2003) is only concerned with the division of cognitive labor, and I note that a concern with the distribution of researcher types may pull in a different direction. This suggests that there is a tradeoff between incentivizing a beneficial division of cognitive labor and a beneficial distribution of researcher types. However, it is difficult to be precise about the nature of this tradeoff until we have a better sense of what is needed to incentivize a beneficial distribution of researcher types.

What is perhaps more interesting is the general question this raises about whether policy proposals designed to address different aspects of the social structure of science are compatible. Strevens (2003) argues that credit should be proportional to impact because he is concerned with the division of cognitive labor. The argument by Strevens (2006) that the Matthew effect is benign also depends on credit being proportional to impact. Heesen (2017) argues that credit should be proportional to the difficulty of the completed task out of a concern with scientists' sharing of intermediate results. The present paper suggests deviating from credit proportional to impact to get the right level of impact-seeking. And other issues may impose yet further, possibly incompatible constraints. Combining and weighting these constraints is an important issue for future research.

5. Conclusion

Lakatos (1978) and Feyerabend (1975) famously argued that a diversity of methodological approaches (what Kitcher, 1990 called a division of cognitive labor) is good for scientific progress. One might reason by analogy that a diversity of researcher types, i.e., of mavericks/impact-seekers and followers/safety-seekers, is also good for scientific progress, as hinted by Kuhn (1996) and Hull (1988) and argued more explicitly by Weisberg and Muldoon (2009) and Thoma (2015). Since credit rewards can arguably be used to incentivize scientists to organize themselves into an optimal division of cognitive labor (Kitcher, 1990; Strevens, 2003), extending the analogy suggests that credit incentives might be used to improve the distribution of researcher types.

At a high level of abstraction, giving out credit for different types of scientific contributions would seem to be an adequate means of incentivizing both a diversity of methodologies and a diversity of researcher types. This makes Thoma's suggestion to use credit for the latter, where Kitcher and Strevens had used it for the former, seem natural. Weisberg and Muldoon (2009) similarly see a close connection there, as evidenced by their use of "division of cognitive labor" in the title of a paper about researcher types. It is somewhat surprising then that as one looks more closely, as I have done here, one finds that the two types of diversity may conflict.

According to Strevens (2003), in order to optimize the division of cognitive labor credit must be equated to impact. I have argued in section 3 that the incentive structure thus created entails a particular distribution of researcher types which is not particularly likely to be a good one. Improving the incentive structure with respect to the distribution of researcher types in the manner suggested in section 4 then requires moving away from the optimum with respect to the division of cognitive labor.

The point generalizes. Recent work on the social structure of science has identified a number of choices scientists make that are potentially influenced by credit incentives. In addition to the division of cognitive labor and the distribution of researcher types, this includes the issues of compensating for the positive externalities of scientific research (Dasgupta & David, 1994; Zollman, 2018), collaboration (Boyer-Kassem & Imbert, 2015), and sharing intermediate results (Heesen, 2017). In each case the good news is that credit rewards can be used to incentivize good or optimal choices.

The bad news identified here is that the right incentive structure may be different for each issue. The division of cognitive labor requires giving credit proportional to impact (per Strevens, 2003); the distribution of researcher types likely requires deviating from this proportionality (per my Theorem 1); incentivizing the sharing of intermediate results requires giving credit proportional to difficulty (per Heesen, 2017); and the other issues may have yet further requirements. It appears that there are real tradeoffs here. A crucial question for future research is how to assess these tradeoffs. In other words, what reward structure is best all-things-considered?

Acknowledgment

Thanks to Kevin Zollman, Michael Strevens, Teddy Seidenfeld, Stephan Hartmann, Adrian Currie, Liam Bright, Shahar Avin, two anonymous referees for the journal, and the audience at the workshop Risk and the Culture of Science in Cambridge for valuable comments and discussion. This work was partially supported by the National Science Foundation under grant SES 1254291 and by an Early Career Fellowship from the Leverhulme Trust and the Isaac Newton Trust.

Appendix A. The tradeoff between impact, success probability, and speed

In this appendix I prove the results from section 3. Recall Assumption 1.

Assumption 1. *The tradeoff function has the following properties:*

1.a. The tradeoff function is decreasing in both of its arguments, i.e., if $p < p'$ then $\lambda(p', c) < \lambda(p, c)$ for any c, and if $c < c'$ then $\lambda(p, c') < \lambda(p, c)$ for any p.

1.b. No perfect work: $\lim_{p \to 1} \lambda(p, 0) = 0$.

ARTICLE IN PRESS

1.c. The tradeoff function is concave: for any $t \in [0,1]$ and for any p, p', c, c'

$$t\lambda(p, c) + (1 - t)\lambda(p', c') \leq \lambda(tp + (1 - t)p', tc + (1 - t)c').$$

I first prove that this assumption entails that the tradeoff function is continuous.

Lemma 1. *If Assumption* 1 *is satisfied,* λ *is continuous (except perhaps when* $p = 1$*).*

Proof. Because λ is concave, it is continuous at any interior point of its domain, i.e., for all $p \in (0,1)$ and for all $c \in (0, \infty)$. It remains to show that λ is continuous on the borders, that is at those points (p, c) with $p = 0$ or $c = 0$. I give a proof for the case $c = 0$ (the other case is similar.)

Fix a value of p. Since λ is decreasing in c, it must be that $\lim_{c\to 0} \lambda(p, c) \leq \lambda(p, 0)$. But it follows from the fact that λ is concave that $\lim_{c\to 0} \lambda(p, c) \geq \lambda(p, 0)$. So the two must be equal, and λ is continuous at the point $(p, 0)$.□

As stated in the lemma, continuity may fail when $p = 1$. But note that, due to concavity and the no perfect work assumption, $\lambda(1,0) \leq 0$. And due to the assumption that λ is decreasing it also follows for any $c > 0$ that $\lambda(1, c) < 0$ and $\lim_{p\to 1}\lambda(p, c) \leq 0$.

Since negative speed has no interpretation in the model, and since the function C obviously will not be maximized at points where the tradeoff function is negative or zero, it makes no difference to the results to work with a "fixed up" function of the tradeoff function that is continuous everywhere (i.e., including at points with $p = 1$). But this does significantly simplify the proofs below. So below I assume that λ is continuous everywhere.

The role of the maximum impact function μ is to identify the set of choices for p and c where the speed is positive.

Proposition 1. (*Maximum Impact*). *Let* $p \in [0,1]$*. Assumption* 1 *entails that there exists a function* μ: $[0,1] \to [0, \infty)$ *such that* $\lambda(p, c) \geq 0$ *if and only if* $c \leq \mu(p)$*. The function* μ *is decreasing and concave.*

Proof. Let $p \in [0,1]$. Since λ is decreasing in p and $\lim_{p\to 1} \lambda(p, 0) = 0$, $\lambda(p, 0) > 0$ if $p < 1$ and $\lambda(1,0) = 0$. Now consider the part of the tradeoff function obtained by holding p fixed and letting c vary. This partial function is maximized when $c = 0$ (because λ is decreasing in c). Due to concavity, it must eventually become negative. Since λ is continuous, by the intermediate value theorem there must be a value of c such that $\lambda(p, c) = 0$. Define $\mu(p)$ to be the smallest value of c such that $\lambda(p, c) = 0$. Because λ is decreasing in c, $\lambda(p, c) > 0$ whenever $c < \mu(p)$ and $\lambda(p, c) < 0$ whenever $c > \mu(p)$. This establishes the existence of the function μ. It remains to show that μ is decreasing and concave.

Let $p < p'$ and suppose for reductio that $\mu(p) \leq \mu(p')$. By definition of μ, $\lambda(p, \mu(p)) = \lambda(p', \mu(p')) = 0$. But since λ is decreasing,

$$\lambda(p, \mu(p)) > \lambda(p', \mu(p)) \geq \lambda(p', \mu(p')).$$

Contradiction. So $\mu(p') < \mu(p)$, i.e., μ is decreasing.

Let p, p', $t \in [0, 1]$. As established above, μ is defined such that

$$\lambda(p, \mu(p)) = \lambda(p', \mu(p')) = \lambda(tp + (1 - t)p', \mu(tp + (1 - t)p')) = 0.$$

Because λ is concave,

$$t\lambda(p, \mu(p)) + (1 - t)\lambda(p', \mu(p')) \leq \lambda(tp + (1 - t)p', t\mu(p) + (1 - t)\mu(p')).$$

Since the left-hand side of the above vanishes,

$$\lambda(tp + (1 - t)p', \mu(tp + (1 - t)p')) \leq \lambda(tp + (1 - t)p', t\mu(p) + (1 - t)\mu(p')).$$

Since λ is decreasing in its second argument, it follows that

$$t\mu(p) + (1 - t)\mu(p') \leq \mu(tp + (1 - t)p'),$$

which establishes concavity.□

Theorem 1. (*Unique Maximum*). *If Assumption* 1 *is satisfied, there exists a unique point* (p^*, c^*) *that maximizes the function C:*

$$C(p^*, c^*) = \max_{p\in[0,1], c\in[0,\infty)} C(p, c).$$

Moreover, $0 < p^* < 1$ *and* $0 < c^* < \mu(p^*)$.

Proof. Recall that the function C is defined for all $p \in [0,1]$ and $c \in [0, \infty)$ by $C(p, c) = cp\lambda(p, c)$. Note first that, as a consequence of Proposition 1, $C(p, c) \geq 0$ if and only if $c \leq \mu(p)$. Since C is a continuous function (as a consequence of λ being continuous), by the extreme value theorem it achieves a maximum at least once on the compact set $\{(p, c)\colon p \in [0,1], c \in [0, \mu(p)]\}$, i.e., there exists at least one point (p^*, c^*) such that

$$C(p^*, c^*) = \max_{p\in[0,1], c\in[0,\mu(p)]} C(p, c).$$

Note further that $C(p, c) > 0$ if and only if $0 < p < 1$ and $0 < c < \mu(p)$. But then certainly it must be the case that $C(p^*, c^*) > 0$. Hence $0 < p^* < 1$ and $0 < c^* < \mu(p^*)$. Moreover, since $C(p, c) < 0$ when $c > \mu(p)$, any point that maximizes C on the restricted domain $\{(p, c)\colon p \in [0,1], c \in [0, \mu(p)]\}$ must also be a global maximum of C:

$$C(p^*, c^*) = \max_{p\in[0,1], c\in[0,\infty)} C(p, c).$$

It remains to show that the maximum is unique. To do this I rely on a theorem of Kantrowitz and Neumann (2005) for products of concave functions.

Let $(p', c') \neq (p^*, c^*)$. To show uniqueness of the maximum, it suffices to show that (p', c') does not maximize C. Since any maximum must satisfy the condition established above, restrict attention to cases with $0 < p' < 1$ and $0 < c' < \mu(p')$.

Let $f\colon [0,1] \to (0, \infty)$ be the function defined by

ARTICLE IN PRESS

$f(t) = C(tp^* + (1-t)p', tc^* + (1-t)c') = (tc^* + (1-t)c')(tp^* + (1-t)p')\lambda(tp^* + (1-t)p', tc^* + (1-t)c')$

for all $t \in [0,1]$. Because C is maximized at (p^*, c^*), f is maximized at $t = 1$.

Note that f is the product of three concave and nonnegative functions: λ is a concave function of t as a consequence of Assumption 1, and $tc^* + (1-t)c'$ and $tp^* + (1-t)p'$ are linear functions of t and hence also concave. Moreover, since either $c^* \neq c'$ or $p^* \neq p'$, at least one of the functions $tc^* + (1-t)c'$ and $tp^* + (1-t)p'$ has a unique maximum (e.g., if $c^* > c'$, $tc^* + (1-t)c'$ is maximized at $t = 1$). Finally, none of the three functions are identically zero on [0,1] (in fact none of the three functions are zero anywhere). So it follows from Kantrowitz and Neumann (2005, theorem 4.iii) that f has a unique maximum at $t = 1$. Hence $C(p', c') = f(0) < f(1) = C(p^*, c^*)$.□

References

Alexander, Jason McKenzie, Himmelreich, Johannes, & Thompson, Christopher (2015). Epistemic landscapes, optimal search, and the division of cognitive labor. *Philosophy of Science, 82*(3), 424–453. https://doi.org/10.1086/681766.

Avin, Shahar. Centralised funding and epistemic exploration. *The British Journal for the Philosophy of Science*, forthcoming. https://doi.org/10.1093/bjps/axx059.

Bornmann, Lutz, & Daniel, Hans-Dieter (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation, 64*(1), 45–80. https://doi.org/10.1108/00220410810844150.

Boyer-Kassem, Thomas, & Imbert, Cyrille (2015). Scientific collaboration: Do two heads need to be more than twice better than one? *Philosophy of Science, 82*(4), 667–688. ISSN 00318248 http://www.jstor.org/stable/10.1086/682940.

Cole, Stephen, & Cole, Jonathan R. (1967). Scientific output and recognition: A study in the operation of the reward system in science. *American Sociological Review, 32*(3), 377–390. ISSN 00031224 http://www.jstor.org/stable/2091085.

Cole, Jonathan R., & Cole, Stephen (1973). *Social stratification in science.* Chicago: The University of Chicago Press ISBN 0226113388.

Dasgupta, Partha, & David, Paul A. (1994). Toward a new economics of science. *Research Policy, 23*(5), 487–521. ISSN 0048-7333 http://www.sciencedirect.com/science/article/pii/0048733394010021.

Feyerabend, Paul (1975). *Against method. New left books.* London.

Grim, Patrick, Singer, Daniel J., Fisher, Steven, Bramson, Aaron, William, J., Berger, et al. (Dec 2013). Scientific networks on data landscapes: Question difficulty, epistemic success, and convergence. *Episteme, 10*(4), 441–464. ISSN 1750-0117 https://doi.org/10.1017/epi.2013.36.

Heesen, Remco (2017). Communism and the incentive to share in science. *Philosophy of Science, 84*(4), 698–716. ISSN 0031-8248 https://doi.org/10.1086/693875.

Huber, John C. (1998a). Invention and inventivity as a special kind of creativity, with implications for general creativity. *Journal of Creative Behavior, 32*(1), 58–72. ISSN 2162-6057 https://doi.org/10.1002/j.2162-6057.1998.tb00806.x.

Huber, John C. (1998b). Invention and inventivity is a random, Poisson process: A potential guide to analysis of general creativity. *Creativity Research Journal, 11*(3), 231–241. https://doi.org/10.1207/s15326934crj1103_3.

Huber, John C. (2001). A new method for analyzing scientific productivity. *Journal of the American Society for Information Science and Technology, 52*(13), 1089–1099. ISSN 1532-2890 https://doi.org/10.1002/asi.1173.

Huber, John C., & Wagner Dobler, Roland (2001b). Scientific production: A statistical analysis of authors in physics. *1800–1900. Scientometrics, 50*(3), 437–453. ISSN 0138-9130 https://doi.org/10.1023/A%3A1010558714879.

Huber, John C., & Wagner-Döbler, Roland (2001a). Scientific production: A statistical analysis of authors in mathematical logic. *Scientometrics, 50*(2), 323–337. ISSN 0138-9130 https://doi.org/10.1023/A%3A1010581925357.

Hull, David L. (1988). *Science as a process: An evolutionary account of the social and conceptual development of science.* Chicago: The University of Chicago Press ISBN 0226360504.

Hull, David L. (1997). What's wrong with invisible-hand explanations? *Philosophy of Science, 64*, S117–S126. https://doi.org/10.1086/392592.

Kantrowitz, Robert, & Neumann, Michael M. (2005). Optimization for products of concave functions. *Rendiconti del Circolo Matematico di Palermo, 54*(2), 291–302. ISSN 0009-725X https://doi.org/10.1007/BF02874642.

Kitcher, Philip (1990). The division of cognitive labor. *The Journal of Philosophy, 87*(1), 5–22. ISSN 0022362X http://www.jstor.org/stable/2026796.

Kuhn, Thomas S. (1996). *The structure of scientific revolutions* (3rd ed.). Chicago: The University of Chicago Press.

Lakatos, Imre (1978). *The methodology of scientific research programmes.* Cambridge: Cambridge University Press.

Leonard, Thomas C. (2002). Reflection on rules in science: An invisible-hand perspective. *Journal of Economic Methodology, 9*(2), 141–168. https://doi.org/10.1080/13501780210137092.

Merton, Robert K. (1957). Priorities in scientific discovery: A chapter in the sociology of science. *American Sociological Review, 22*(6), . (ISSN 00031224 http://www.jstor.org/stable/2089193 (Reprinted in Merton (1973, chapter 14) ()).

Merton, Robert K. (1969). Behavior patterns of scientists. *American Scholar, 38*(2), . (ISSN 00030937 http://www.jstor.org/stable/41209646 (Reprinted in Merton (1973, chapter 15) ()).

Merton, Robert K. (1973). *The sociology of science: Theoretical and empirical investigations.* Chicago: The University of Chicago Press ISBN 0226520919.

Strevens, Michael (2003). The role of the priority rule in science. *The Journal of Philosophy, 100*(2), 55–79. ISSN 0022362X http://www.jstor.org/stable/3655792.

Strevens, Michael (2006). The role of the Matthew effect in science. *Studies in History and Philosophy of Science, 37*(2), 159–170. ISSN 0039-3681 http://www.sciencedirect.com/science/article/pii/S0039368106000252.

Pöyhönen, Samuli (2017). Value of cognitive diversity in science. *Synthese, 194*(11), 4519–4540. ISSN 1573-0964 https://doi.org/10.1007/s11229-016-1147-4.

Sinatra, Roberta, Wang, Dashun, Deville, Pierre, Song, Chaoming, & Barabási, Albert-László (2016). Quantifying the evolution of individual scientific impact. *Science, 354*(6312), aaf5239. ISSN 0036-8075 http://science.sciencemag.org/content/354/6312/aaf5239.

Thoma, Johanna (2015). The epistemic division of labor revisited. *Philosophy of Science, 82*(3), 454–472. ISSN 00318248 http://www.jstor.org/stable/10.1086/681768.

Weisberg, Michael, & Ryan, Muldoon (2009). Epistemic landscapes and the division of cognitive labor. *Philosophy of Science, 76*(2), 225–252. ISSN 00318248 http://www.jstor.org/stable/10.1086/644786.

Wray, K. Brad (2003). Is science really a young man's game? *Social Studies of Science, 33*(1), 137–149. ISSN 03063127 http://www.jstor.org/stable/3183139.

Zollman, Kevin J. S. (2018). The credit economy and the economic rationality of science. *The Journal of Philosophy, 115*(1), 5–33. https://doi.org/10.5840/jphil201811511.