# To Expedite the Flow of Knowledge

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy

Department of Philosophy
Carnegie Mellon University

Remco Heesen

Pittsburgh, PA
May 13, 2016

# To Expedite the Flow of Knowledge

Remco Heesen

## Abstract

This dissertation focuses on the epistemic consequences of scientists' decisions regarding journal publications. Chapters 2 and 3 consider the author's perspective, investigating what decisions would be made by scientists aiming to maximize the credit they receive for their work. Here, credit is awarded according to the priority rule, which says that recognition for scientific work depends on its originality. Chapter 4 considers the editor's perspective, where the editor's interest is taken to be in the credit accrued to papers she accepts for publication. In each case the goal is to normatively evaluate the incentive structure of science.

"Communism and the Incentive to Share in Science" (chapter 2) identifies a positive effect of the priority rule. It encourages scientists to share their work rather than keep it secret, even when keeping it secret would increase their chances of future publications. Thus the priority rule can explain the origins and persistence of the communist norm, the institutional norm which mandates that scientists share their results. Because other scientists can build on work that is shared, this is a positive effect.

In contrast, "Expediting the Flow of Knowledge Versus Rushing into Print" (chapter 3) shows that the priority rule may lead to "rushing into print" behavior, where scientists publish their results prematurely. While scientific work can never be guaranteed to be error-free, I argue that the priority rule incentivizes scientists to err too much on the side of speed in

the tradeoff between speed and accuracy. As a result more errors enter the scientific literature than would be epistemically optimal.

In "When Journal Editors Play Favorites" (chapter 4) I consider two arguments in favor of triple-blind reviewing. The first argument claims that editors who fail to practice triple-blind review hurt authors. I endorse this argument and argue more specifically that epistemic injustices are committed against authors. The second argument claims that editors who fail to practice triple-blind review hurt their readers, as their biases influence the quality of the papers they accept. While biases may have a negative effect on quality, I show that revealing identity information to the editor may also have a positive effect on quality by reducing uncertainty. The net effect is unclear, except in certain scientific fields where I argue the positive effect does not apply. As a result, I endorse the second argument only for those fields.

An interesting theme that emerges from both chapters 2 and 3 is the idea that there is not one unique priority rule. The priority rule can be implemented in different ways, depending on what kind of scientific contributions are given credit, and how much credit is given for them. In the conclusion (chapter 5) I raise some questions for future research about the epistemic consequences of these different ways of implementing the reward structure of science.

# Acknowledgements

While working on this dissertation, I have benefited from the intellectual input and emotional support of many individuals. I would like to use this space to thank some of these people.

First and foremost, I want to thank my adviser, Kevin Zollman. I first became interested in studying social epistemology when I saw him give a talk in 2011. This talk used computer simulations to study the question whose testimony to trust (for those who are interested, the paper is published as Zollman 2015). His intellectual influence on the present document will be unmistakable to all who are familiar with his work. Moreover, Kevin has been a very supportive adviser through five years in Pittsburgh. He has always been able to find time when I needed advice, and perhaps most important of all, I always left our meetings feeling good about my work and with a clear idea of what the next steps were.

Thanks also to the other members of the dissertation committee: Michael Strevens, Stephan Hartmann, and Teddy Seidenfeld. They have enthusiastically supported this dissertation from the beginning while providing critical remarks when needed. This work is much better than it would otherwise have been thanks to their help.

I have been fortunate to have been in Pittsburgh at a time when the social epistemology of science was a topic of interest for a number of other graduate students. Being able to discuss relevant questions on a regular basis with other experts—sometimes in a formal reading group, sometimes

more informally—has been extremely helpful, as well as a lot of fun. Thank you, Liam Kofi Bright, Haixin Dang, Taku Iwatsuki, Zina Ward, and other attendees.

Liam deserves special mention. We share not only much of our research interests, but also have very similar values about how that research should be done and how it should be presented. Liam has read more of my work than anyone else, and has given me lots of detailed and helpful feedback. My favorite is when he points out gratuitous words or turns of phrase (e.g., "clearly", "it is obvious that"), which we both strongly disapprove of, but find ourselves using involuntarily.

In addition to those mentioned above, I have received valuable comments and suggestions from Thomas Boyer-Kassem, Lee Elkin, Cailin O'Connor, and Jan Sprenger. I also wish to thank the numerous people who asked helpful questions when I presented parts of this work at the Conference in Formal Epistemology in Bristol, at the Congress of Logic, Methodology, and Philosophy of Science and the Logic Colloquium in Helsinki, at the "Show 'n Tell" graduate student seminar in Pittsburgh, and at invited talks in Singapore and Tilburg.

The honor of being the only person acknowledged here whom I have never actually met goes to Steve Fuller. I was reading his book *Social Epistemology* in a park in San Francisco in March 2014 when I came across the following line:

> [T]he scientist is supposed to *both* expedite the flow of knowledge *and* not rush into print. But how can he "expedite" without also "rushing"? (Fuller 2002, p. 201)

In the context, this is clearly a rhetorical question, but I made it my goal to answer it. Chapters 2 and 3 are a direct result of my attempt to do so. I have acknowledged this debt in the title of the dissertation and the title of chapter 3.

Thanks to the National Science Foundation, which funded some of the research that went into this dissertation through grant SES 1254921. The usual disclaimer applies.

On a more personal level, I have had the good fortune of being close friends with many of my fellow graduate students here in Pittsburgh, and even (gasp!) some people of a non-academic persuasion. I would like to thank them for distracting me at suitable times.

I also want to thank my Dutch friends: for being eager to hang out whenever possible, and for making sure a visit to the Netherlands still feels like coming home.

I want to thank my girlfriend Becky for her love and support as I combined being on the job market with finishing this dissertation, while she herself was busy finishing up her master's degree.

And finally I thank my parents, Frank and Sylvia. Both their unconditional support and sage advice have been of immeasurable value to me in navigating life, regardless of whether we were living in the same house or half a world apart.

Thank you all!

# Contents

# Chapter 1

# Introduction

## 1.1 The Social Epistemology of Science

Philosophy of science has existed as a separate sub-discipline of philosophy, with its own journals, conferences, and experts, for about 150 years. Although of course the questions philosophers of science ask had been raised by other thinkers long before that. Aristotle and Francis Bacon are two of the most prominent among these predecessors.

Some of the questions philosophers have science have been concerned with include the following. Is there a unique theory that explains (or unifies, or predicts) a given body of evidence (Duhem 1906, Quine 1951)? Why do we accept or reject certain theories (Popper 1959)? What reasons might we have for introducing new entities into our theories (Carnap 1950, Hempel 1958)? Do scientific theories aspire to (objective) truth and do the entities posited by them exist (van Fraassen 1980, Boyd 1983)?

These questions are largely focused on the content of science. They may be unified under a question that reads something like: What would an epistemically rational scientist believe? This unification suggests a particular perspective on the above questions, and the history of philosophy of science shows that these questions have often been fruitfully addressed from this

perspective.

But in taking that perspective, a number of features of science are ignored. I want to focus attention on two of these features here.[1] First, science is practiced by a large number of scientists rather than by one individual. Second, individual scientists are not epistemically rational.

Taking these features of science seriously may involve one of two things. First, it suggests a different perspective from which to address some of the traditional questions of philosophy of science. Take for example the question which theories scientists should accept. If approached from the perspective of a single epistemically rational scientist, the goal is to identify features of scientific theories that make them more or less acceptable, with the hope that a uniquely most acceptable theory can be identified in some or all contexts.

When the question is approached from the perspective of a group of scientists whose goals and level of rationality may differ, that interpretation of the question retains its relevance, but new interpretations are suggested as well. For example, if there is a uniquely most acceptable theory, should all scientists accept it or is it better if some diversity is retained? A number of philosophers have argued the latter (Feyerabend 1975, Kitcher 1990, Zollman 2010). And supposing the epistemically most beneficial distribution of scientists among theories was known, what kind of goals or decision rules that individual scientists might have would lead them to actually distribute themselves in this way (Strevens 2003, Mayo-Wilson et al. 2013)?

So taking seriously the facts that science is not done by an individual and that individual scientists are not epistemically rational suggests new approaches to old questions. But, secondly, taking these facts seriously may also raise new philosophical questions. Consider the following examples.

[1] I am not, of course, suggesting that those philosophers of science who are interested in what an ideally rational scientist would do are unaware of these features of science. Rather, I take it they would argue that an answer to the question they are interested in retains its (normative) relevance to science despite these features. I view this kind of work as complementary to, rather than in competition with, the kind of work I discuss below.

1. Which information (e.g., results of scientific experiments) should be shared and distributed (Bala and Goyal 1998, Zollman 2009, 2010, Boyer 2014)?

2. How (if at all) does erroneous or fraudulent work disrupt science, and what can be done about this (Bruner 2013)?

3. How should scientific work be rewarded (Dasgupta and David 1994, Strevens 2003)?

These questions do not even arise if science is studied as if it was the work of a single epistemically rational scientist, but they are natural questions from this alternative perspective. Both new approaches to old questions and new questions are of interest to those philosophers who ask normative questions regarding the social structure of science. I will refer to this field of study as *the social epistemology of science*.

At least until recently, the questions studied by the social epistemology of science have not received as much attention from philosophers of science. However, these questions are equally venerable: they can trace their roots back at least as far as Bacon.

In his book *New Atlantis*, Bacon (1626) describes a utopian society in which science plays an important role. The scientists are organized in "Salomon's House", an anticipation of the modern university and purportedly a model for the Royal Society. The leader of Salomon's House explains how science is practiced there, and he has something to say about each of the above three questions.

Regarding the distributing of information, he says:

> [W]e have consultations, which of the inventions and experiences which we have discovered shall be published, and which not; and take all an oath of secrecy for the concealing of those which we think fit to keep secret; though some of those we do reveal sometime to the State, and some not. (Bacon 1626)

So here all information is shared among the scientists, but not necessarily communicated to the state or the public. In order to facilitate information sharing among scientists, as well as in order to minimize disruptions due to errors or fraud, a high value is placed on honesty:

> [W]e do hate all impostures and lies, insomuch as we have severely forbidden it to all our fellows, under pain of ignominy and fines, that they do not show any natural work or thing adorned or swelling, but only pure as it is, and without all affectation of strangeness. (Bacon 1626)

The threat of ignominy is here used as a way to encourage scientists to follow the norms. This is part of a wider pattern in which honor and dishonor are used to reward and punish good and bad scientific work, respectively (in addition to monetary rewards and fines):

> [U]pon every invention of value we erect a statue to the inventor, and give him a liberal and honorable reward. (Bacon 1626)

In *Novum Organon*, Bacon (1620 [1858]) laments the institutions of his time for not being conducive to good scientific work:

> [I]n the customs and institutions of schools, academies, colleges, and similar bodies destined for the abode of learned men and the cultivation of learning, everything is found adverse to the progress of science. (Bacon 1620 [1858], Book I, aph. 90)

At least part of the problem, according to Bacon, is that the right system of rewards for scientific work is not in place:

> [I]t is enough to check the growth of science, that efforts and labours in this field go unrewarded. For it does not rest with the same persons to cultivate sciences and to reward them. The

> growth of them comes from great wits; the prizes and rewards of them are in the hands of the people, or of great persons, who are but in very few cases even moderately learned. (. . . ) And it is nothing strange if a thing not held in honour does not prosper. (Bacon 1620 [1858], Book I, aph. 91)

In this passage Bacon argues not only that successful scientific work should be rewarded with honor (as well as money), but that (the distribution of) these rewards should be determined by scientists themselves. So here is an argument not only for giving credit for scientific discoveries, but also for *peer review*.

## 1.2 The Role of Journals in Science

One area where the questions I raised above naturally come together is that of *scientific journals*. Journals' main purpose is information sharing (question 1); they aim to minimize the spread of erroneous or fraudulent results through the system of peer review, but when an error does get published in a journal it may help that error persist (question 2); and getting published in a journal is itself part of the rewards given for successful scientific work (question 3).

Journals act as the clearing houses of scientific information. If a scientific result is not published, few scientists are likely to be aware of it. Given their key role in the transmission of information in science, they should be a central area of interest for the social epistemology of science.

The system of peer review is of particular interest. Its goals are explicitly epistemic: to accept "good" scientific work for publication and reject "bad" scientific work. Some of the questions one might ask about the role of journals include:

1. Why do scientists choose to publish their work in a journal?

2. How much time should scientists spend on a given research project before publishing their results, and how can they be incentivized to spend no more and no less time?

3. How do various biases enter the system of peer review and what ethical and epistemic problems does this raise?

4. Should scientific journals be abolished in favor of some other system of transmitting information between scientists?

The first three of these questions will be addressed in this dissertation; they each motivate one of the central chapters. The fourth question I raise only to set it aside. Journals and peer review have many flaws: they are slow, the way they evaluate publications is unreliable, and it involves various biases (see Smith 2006, Lee et al. 2013, and references therein). Especially given the rise of the Internet, it is not clear that journals are still the best way to transmit information. Various alternatives have been proposed, such as post-publication peer review.

For now, however, journals remain the dominant way in which scientists publicize their work. My discussion in this dissertation hence focuses on journals and peer review as they currently exist. However, this is not to say that this dissertation becomes obsolete if journals were to be abolished. The argument of chapter 2 relies only on the assumption that there is a distinction between sharing scientific work and keeping it secret, and hence is independent of any assumption about the existence of journals or the way peer review works. The argument of chapter 3 relies on the existence of peer review only insofar as it entails that the rewards for scientific work depend in a significant way on a short-term evaluation of its long-term impact (which I argue in the chapter to be almost inevitable). Chapter 4 does assume something like the current system of peer review, as it identifies some of the biases in that system and analyzes their effects.

I am not the first to recognize the importance of journals. The remainder of this section gives an overview of the work that already exists.

Let me start on the empirical side. In the sociology and the economics literature there exists a large range of empirical work analyzing journals, peer review, scientists' decisions to publish, errors, fraud, and retractions in journal publications, and so on.

A good place to start is Price (1963, 1965), who observed that the total volume of scientific journal publications increases exponentially, and that the distribution of journal publications per author is highly skewed as well: many if not most scientists produce only a single paper, but a small handful produce dozens or even hundreds. Cole and Cole (1973) confirm these observations and show that the number of publications a scientist produces correlates highly with other measures of prestige, such as the number of citations to a scientist's work, the prestige of their institution, and the receipt of prestigious awards.

Merton (1942) identifies a number of *institutional norms of science*: behavioral rules that scientists tend to follow and which appear to have normative force (in the sense that scientists perceive a duty to conform to them). These are universalism, communism, disinterestedness, and organized skepticism. Recent empirical work has tested the support for these norms among contemporary scientists, which ranges from middling to very strong (Macfarlane and Cheng 2008, Anderson et al. 2010). Of these norms, communism, which requires that scientists widely share the results of their work, is the most directly relevant to journal publications. I analyze it in detail in chapter 2.

Philosophers like Quine (1951), Kuhn (1962), Lakatos (1968), and Feyerabend (1975) have argued that any scientific claim is in principle open to revision. One way this plays out is in journal publications that contradict previously published scientific work. Sometimes these disproven publications are retracted, although usually mere contradiction is not enough for this:

only serious flaws, such as suspected fraud, tend to lead to retraction (Budd et al. 1998). Somewhat worryingly, even publications that have been retracted and/or thoroughly proven to be erroneous continue to be cited as if they were correct (Budd et al. 1998, Tatsioni et al. 2007).

In some sciences, such as medicine and psychology, many contributions depend on drawing conclusions from large data sets using statistics. In recent years worries have been raised about the *reproducibility* of such work. Two relatively small studies in medicine found that a majority of results fail to reproduce (Prinz et al. 2011, Begley and Ellis 2012), and a large systematic study found the same in psychology (Open Science Collaboration 2015).

Some theoretical work has aimed to explain this problem of reproducibility. One contributing factor is the so-called file drawer problem—studies that yield negative results are not published because they are considered uninteresting—which creates a bias towards positive results in the literature (Rosenthal 1979, Ioannidis 2005). A related factor is "researcher degrees of freedom"—the various choices that go into designing and analyzing a study—which similarly create a bias towards positive results (Ioannidis 2005, Simmons et al. 2011).

For my part, I consider empirical results that fail to reproduce to be a special case of scientific results that turn out to be erroneous. While the work cited in the previous paragraph helps explain what aspects of scientific practice are responsible for relatively high rates of error, my aim in chapter 3 is to show why it is rational for scientists, given the incentive structure presented to them, to engage in practices that would lead to high rates of error.

A lot of research exists on peer review. Here I distinguish between subjective research, which aims to measure scientists' perception of peer review, and objective research, which aims to measure aspects of peer review directly.

On the subjective side, Bailey et al. (2008a,b) find that scientists generally feel quite positive about the system of peer review, judging it as mostly

fair and unbiased (in surveys in the fields of accounting and finance). However, scientists feel less positive about the time it takes for a paper to be reviewed, and they worry about editorial favoritism (more on the latter below). Ziobrowski and Gibler (2000) ask scientists which factors they consider in choosing a journal to submit their work to (in a survey in the field of real estate). The most important factor is the perceived quality of the journal, but promotion and tenure considerations and the ease and perceived fairness of the editorial process also play a role.

On the objective side, a key study is Blank (1991), who ran a randomized controlled trial at *The American Economic Review* in which some submissions are reviewed single-blind and others double-blind. Under single-blind review, the reviewer's name is not revealed to the author, but the reviewer knows the name of the author, whereas under double-blind review the reviewer also does not know who the author is. Blank finds that reviewers are more critical when double-blind review is used. She also finds that female authors benefit from double-blind review, suggesting some bias against women, but not significantly so.

Building on this work, Laband and Piette (1994b) find that double-blind review is more effective at identifying high-quality papers (if quality is measured by future citations), but journals that practice single-blind review attract better submissions and end up outperforming journals that practice double-blind review in total citations. Budden et al. (2008) present evidence that switching to double-blind review increases the percentage of publications by female authors.

The idea that journal editors and reviewers may display biases is not new (Crane 1967). A study by Wennerås and Wold (1997) provides evidence for two kinds of biases: favoritism (inflating the evaluation when the editor or reviewer knows the scientist whose work is being evaluated) and gender bias (deflating the evaluation of work by women). Lee et al. (2013) survey both empirical and normative work on biases in peer review. Here I give a short

survey of empirical work on favoritism and gender bias.

Editorial favoritism happens when papers written by authors whom the editor knows are accepted for publication at a higher rate than papers written by authors unknown to the editor. Unfortunately, acceptance rates can only be measured if one has inside access to the peer review system, so that one can count rejected papers and keep track of author characteristics for those papers. The study by Blank (1991) appears to have been unique in having this kind of access, but she does not include a variable marking whether or not the editor knows the author in her study.

While direct evidence for favoritism thus does not exist, some strong indirect evidence is available. Laband (1985) and Piette and Ross (1992) measure how many journal pages are allocated to each paper and use this as a proxy for the editor's level of support for the paper. They find that editors allocate more pages per accepted article when they know the author than when they do not know the author.

Editorial favoritism has widely been taken to be a bad thing (Bailey et al. 2008a,b). However, if the editor's duty is taken to be to publish those papers of maximum citation impact, the surprising result is that this disapproval is mistaken. When the distinction between authors known to the editor and authors unknown the editor is made, it has consistently turned out that papers by authors in the former group are more highly cited than papers by authors in the latter group (Laband and Piette 1994a, Smith and Dombrowski 1998, Medoff 2003).

With regard to gender bias, there have been several studies in which academics were presented with job applications that differed only in whether there was a stereotypically male or a stereotypically female name at the top (Steinpreis et al. 1999, Moss-Racusin et al. 2012). Both male and female academics rated the application from a purportedly male candidate more highly than the same application from a purportedly female candidate.

These studies as well as those by Blank (1991), Wennerås and Wold

(1997), and others suggest that there is (implicit) bias against women in various academic contexts including peer review. This evidence has been disputed by some who say that "some of these claims are no longer valid" (Ceci and Williams 2011, p. 3157), but see Lee (forthcoming) for further discussion.

The work cited above is almost exclusively of a descriptive nature. This is of course not a criticism; this work is very valuable and I return to it repeatedly throughout this dissertation. But it does not address normative questions such as: When is it rational for a scientist to publish in a journal? Or: How many erroneous results in a journal is too much? Or: What are the epistemic and ethical effects of biases in peer review? Thus, I finish this section with an overview of more normatively oriented work on the role of journals in science (necessarily brief as not nearly as much exists), before turning to my own contribution.

On the question of what and when to share in journals, Bala and Goyal (1998) and Zollman (2009, 2010) have explored models in which limiting the information available to individual scientists helps a scientific community maintain a kind of diversity that promotes long-run convergence to the truth. The literature on informational cascades uses a different kind of models but arrives at similar conclusions (Banerjee 1992, Bikhchandani et al. 1992, Gale and Kariv 2003).

Boyer (2014), on the other hand, shows that when multiple scientists are working on a problem which can be split up into discrete stages, it is both in their individual interest and in the interest of the group if they publish any progress they make. Strevens (forthcoming) and Banerjee et al. (2014) makes similar claims. I discuss the work of these authors in more detail in chapter 2.

Bruner (2013) uses tools from (evolutionary) game theory to show how rewards for scientists who catch errors can be used to keep scientists honest. The work of Code (1991) and Fricker (2007), while not explicitly focused on

journals, explores the epistemic and ethical consequences of gender bias.

Ellison (2002a) provides a model that attempts to explain the phenomena observed in his own empirical work (Ellison 2002b), which identifies a "slowdown" in the peer review system in economics. In this model, authors need to split their time between improving the original ideas in their paper and improving other aspects of the quality of their paper (explicating the main ideas clearly, embedding them in the literature, testing for robustness, and so on). The main focus of the model is on the standards used by referees and editors in evaluating papers and how those might evolve over time.

Besancenot et al. (2012) attempt to show that more demanding editors can improve the quality of their journal. Faria (2005) models a game between authors and editors. Here, the main result is that authors' impatience helps, while editors' impatience detracts from, the journal's ability to be selective, and hence the quality of the papers it publishes.

Heintzelman and Nocetti (2009) raise the question which journal(s) it is best for scientists to submit their papers to, and in what order. Their results largely vindicate the old adage: try your luck with the most highly regarded journals first, then work your way down. They also provide some evidence that journals use long submission delays to discourage authors. While this may improve the quality of the journal, the authors suggest that using submission fees instead of delays might be more effective at this, and would be an improvement for both authors and editors.

Having outlined some of the most important extant work on the role of journals in science, I now turn to describing my own approach.

## 1.3 The Incentive Structure of Science

As I emphasized in the previous section, scientific journals play an important role in transmitting scientific information. But there is another way in which journals are important to the social structure of science. Namely, journals

play a key role in the *reward structure of science*.

The primary way in which scientific achievements are rewarded is through prestige (Merton 1957).[2] Two important ways to gain prestige are to publish scientific achievements (often, but not always, in a journal) and to have one's publications cited (often, but not always, in another journal publication). Counting journal publications and citations turn out to be excellent ways to measure prestige in science (Price 1965, Cole and Cole 1973).

Not only are scientists rewarded with prestige, but so in turn are journals. The prestige hierarchy of journals is determined by such factors as how many times papers in it have been cited, the acceptance rate of the journal, the individual prestige of the editor, and so on. Publishing in more prestigious journals and getting cited adds to the prestige of the scientist, and having prestigious scientists publish in it and get cited adds to the prestige of the journal.

Rich rewards await the most prestigious scientists. They get to work at the most prestigious universities, where they have big offices and many graduate students. They may win awards and be named to exclusive societies. And they may have something they discovered named after them.

In order to become a prestigious scientist, a scientist must receive *credit* for her work. Receiving credit is here meant in a dual sense. The scientist wants credit for her work, i.e., she wants it to be known that it was her who made this particular contribution. And the scientist wants the work to be credited, i.e., she wants the content of the contribution to be recognized as important. A journal publication is a way to establish the first kind of credit, and is generally seen as necessary, if not sufficient, to achieve the second kind of credit.

[2]Of course scientists are also compensated financially. But not nearly to the same extent as, e.g., in business, both in absolute terms (one can make more money in the private sector than in the academic world) and in relative terms (the differences in salaries are much larger in business than they are among scientists). The prevailing attitude among scientists appears to be that they are not in it for the money.

Scientists are thus in a competition for credit. And the consequences of being successful (or not) can be severe. Successful scientists get the best jobs, win large research grants and prizes, and may even achieve eponymous fame (Merton 1957). Unsuccessful scientists may find themselves without a job, or with a job that provides them few resources for research.

As a result, anyone who wishes to build a career in academic science must care about credit. This point has been made by various philosophers and sociologists of science, e.g., Hull (1988, chapter 8), Kitcher (1990, 1993, chapter 8), Strevens (2003), Merton (1957, 1969), and Latour and Woolgar (1986, chapter 5). As far as I can see, the concern for credit is the only interest that all scientists have in common. Any other motivations, e.g., to advance human knowledge, are idiosyncratic and may not be shared by all scientists.

Throughout most[3] of this dissertation, I assume that scientists are *rational (expected) credit-maximizers*.[4] I make this assumption not because I think scientists have no other motivations besides their concern for credit; they clearly do. Rather, my purpose is to study what scientists are motivated to do *purely as a consequence of their concern for credit*, i.e., in isolation of their other goals.

Given this assumption, the kinds of conclusions my analyses yield are not of the form "scientists should be expected to act in the following way". Instead, conclusions take the form "scientists have a credit incentive to act in the following way". Based on whether or not the way scientists have a credit incentive to act seems praiseworthy or not, this allows me to normatively appraise *the incentive structure of science*, regardless of whether real scientists follow these incentives or not.

[3] I briefly drop the assumption in section 2.7, for reasons to be explained in that section.

[4] In chapter 4, where the focus is on the role of the journal editor rather than on the scientists writing scientific papers, I use an appropriately modified version of this assumption: I assume that the editor accepts all and only those papers that have a high expected citation impact.

So when I speak of what rational credit-maximizing scientists would do, it is not my aim to judge real scientists (even if they do not act like rational credit-maximizers, they may well be rational relative to their complete set of goals). My aim is rather to judge the incentive structure of science.

I am not the first to study the incentive structure of science by looking at what rational credit-maximizing scientists would do, but the list of previous work is quite small. It includes Kitcher (1990, 1993, chapter 8), Dasgupta and David (1994), Strevens (2003, forthcoming), Bright (forthcoming), Boyer (2014), and Boyer-Kassem and Imbert (2015). This dissertation can be added to that list. Let me now briefly introduce the three chapters that form the main body of this work.

In chapter 2, the main question is whether scientists have a credit incentive to follow the communist norm, which requires the results of scientific work to be widely shared. Strevens (forthcoming) has argued that such an incentive does not exist in general. I present a model in which such an incentive does exist, and I argue the model applies to most research situations a scientist may find herself in.

The question in chapter 3 is whether scientists' credit incentive to trade off speed against reliability (or speed against reliability and impact, see section 3.4) aligns with the socially optimal way to make that tradeoff. I present a model in which there is a structural misalignment between the credit-maximizing behavior and the socially optimal behavior. I argue that this model captures the situation most scientists are in most of the time.

Chapter 4 considers the role of the journal editor. I focus on the question whether the editor should be informed of the identity of the authors of submitted papers, i.e., whether the journal should practice triple-blind review. I argue that failing to practice triple-blind review harms the authors of submitted papers. But there is not necessarily a corresponding harm to the readers of the journal: whether the average quality of accepted papers is increased or decreased by switching to triple-blind review depends on the

context. Only in some fields (e.g., mathematics and some humanities) is triple-blind review always preferable from both perspectives.

## 1.4 The Priority Rule and Formal Models

To conclude this introductory chapter, I indicate a number of themes that recur throughout this dissertation. Two of these have already been mentioned: the subject matter (the role of journals in science) and the kinds of questions I ask (i.e., questions about the incentive structure of science). Here I add three more: the priority rule as a way of making scientists' credit incentives more precise, the use of formal models as a methodological tool, and the practical implications of this kind of work.

Throughout this dissertation, I ask the question: what would rational credit-maximizing scientists do? The answer to this question depends, among other things, on how credit is allocated. This is where the *priority rule* comes in. First formulated by Merton (1957), I use Strevens' statement of the rule:

> [R]ewards to scientists are allocated solely on the basis of actual achievement, rather than, for example, on the basis of effort or talent invested [and] no discovery of a fact or procedure but the first counts as an actual achievement. (Strevens 2003, p. 56)

Because of this role of priority in determining who gets credit for scientific work, scientists spend a lot of time and energy on *priority races* (trying to beat their colleagues to a particular result, see Strevens 2003) and *priority disputes* (arguing over who was the first to achieve a particular result, see Merton 1957).

The priority rule is directly related to journal publications. In order to be seen to be the first to achieve a particular result, a scientist needs to publish her work. So publishing is crucial to getting credit for her work, which is in turn crucial for building her career (see section 1.3). This helps

explain scientists' eagerness to publish their work. I return to this eagerness to publish in each chapter of this dissertation.

The three chapters that form the body of this dissertation each use the priority rule to illuminate some aspect of the social structure of science. But, as a whole, these chapters also illuminate the priority rule itself. In particular, it turns out that there is no such thing as *the* priority rule. There are in fact multiple versions of the priority rule: depending on what counts as a "discovery of a fact or procedure" and how much credit is given for a particular discovery, different priority rules and hence different incentive structures are created. I argue this point in more detail in the conclusion of this dissertation (see section 5.2).

A clarificatory note: I use the word "science" (and "scientist") in a broad sense, like German *Wissenschaft* or Dutch *wetenschap*. This includes (in additional to the natural sciences) the social sciences and the humanities. While I do not emphasize this, some of my results may even extend beyond the academic world to other fields where the priority rule applies, such as art or journalism.

Throughout this dissertation, I combine philosophical reasoning with mathematical models to answer questions about what rational credit-maximizing scientists would do in particular situations. The dissertation is therefore an example of *mathematical philosophy* (Leitgeb 2013). I discuss some of the virtues of this approach in section 5.3. Here, I introduce some of the formal modeling techniques that I use repeatedly in what follows.

Throughout, I draw on rational choice theory, or more specifically *decision and game theory*. In order to determine the behavior of a rational credit-maximizing scientist, I assume such a scientist is appropriately modeled as a Bayesian subjective expected utility maximizer, where the utility function is just a measure of how much credit the scientist acquires. While I am aware of various challenges to the subjective expected utility model (some of my previous work has focused on this, see Sprenger and Heesen 2011), I assume

that for the cases considered in this dissertation this model captures what a rational agent would do, or comes close enough so as not to make a difference.

The priority rule helps determine the shape of the utility function. Credit is acquired if and only if a contribution is made, and the contribution is published before another scientist publishes it. An interesting feature of the priority rule helps narrow down what this means. Merton attributes this feature to François Arago, permanent secretary of the French Academy of Sciences in the nineteenth century, whom he quotes as follows:

> "About the same time" proves nothing; questions as to priority may depend on weeks, on days, on hours, on minutes. (Merton 1957, p. 658)

This is arguably the harshest part of the priority rule, and Merton disapproves of it for this reason.[5] But harsh or not, this appears to be how the priority rule is in fact used, and from a modeler's perspective this is actually quite convenient. The priority rule does not compromise: credit always goes to the first scientist to make the contribution, no matter how close behind the next scientist is.

In order to properly represent this aspect of the priority rule, a continuous-time model is needed. A model with discrete time units will not do as it may place two contributions in the same time unit even though they were not exactly simultaneous.

A scientist in this model has expectations about when she will finish the project she is working on. These are reflected in a (subjective) probability distribution. I assume that she uses the *exponential distribution* (a detailed defense of this assumption is given in section 2.4). This assumption allows

[5] Merton (1957, pp. 658–659) argues that it represents a pathological extreme: when the interval between two discoveries is so small, "priority has lost all functional significance." I agree with Strevens (2003, section IV.1) that this is not obviously correct. A version of the priority rule which gives shared credit when the time interval between discoveries is below a certain threshold would create a different incentive structure for scientists, and it is an open question whether that incentive structure would be better or worse.

for scientists to work at different speeds by changing the parameter $\lambda$, which gives the average productivity of the scientist.

The combination of the priority rule with exponential distributions makes for an elegant model with a number of mathematically convenient features, which I discuss in chapter 2 (see especially section 2.4). In contrast, neither the priority rule nor exponential distributions are mentioned explicitly in chapter 3. However, the model given there fits nicely with this framework, as I explain next.

Chapter 3 emphasizes that scientists can choose to some extent how fast to work, by deciding how much work to do before publishing the results. This is reflected in the choice of a parameter $\lambda$, which gives the average speed of a scientist. The (implicit) underlying idea is that the scientist cannot control exactly how long a particular project will take (this is an exponentially distributed random variable), but she can manipulate her *average* speed by setting a standard of reliability for her own work.

If the parameter $\lambda$ of chapter 3 is interpreted as the parameter of an underlying exponential distribution, the priority rule is assumed to apply, and if the scientist immediately starts a new project whenever another scientist publishes results that pre-empt the project she was working on, then mathematical features of the exponential distribution guarantee that the scientist's expected credit per unit time is as given in chapter 3 regardless of whether or not other scientists are ever working on the same project as her. Hence, under these assumptions, the strategic aspects of credit-maximization (those that require game theory) can conveniently be ignored and a decision-theoretic analysis can be given, which is exactly what I do.

Finally, I want to emphasize that the investigation of the incentive structure of science carried out here is not just of philosophical interest. Whenever I give a normative appraisal of the existing incentive structure, and in particular when I compare it with a hypothetical alternative, I effectively give a recommendation of what the incentive structure of science should be.

No one party controls the incentive structure of science: it is shaped by the actions of all scientists and other involved groups and agencies. But science policy makers (in governments or other agencies) can have significant influence on this incentive structure. My normative appraisals are, in part, recommendations to them. I draw out these aspects of my conclusions in more detail in section 5.4.

# Chapter 2

# Communism and the Incentive to Share in Science

## 2.1 Introduction

The social value of scientific work is highest when it is widely shared. Work that is shared can be built upon by other scientists, and utilized in the wider society. Work that is not shared can only be built upon or utilized by the original discoverer, and would have to be duplicated by others before they can use it, leading to inefficient double work.[1]

To put the point more strongly, work that is not widely shared is not really scientific work. Insofar as science is essentially a social enterprise, representing the cumulative stock of human knowledge, work that other scientists do not know about and cannot build upon is not science (cf. the distinction between Science and Technology in Dasgupta and David 1994). The sharing of scientific work is thus a necessary condition not merely for the success of science, but in an important sense for its very existence.

[1]Of course scientific work is often duplicated by others even when it is shared (so-called replications). But this is not inefficient in the same way, as after the replication is shared the work is known by all to be more certainly established than if only one or the other instance was shared.

The sociologist Robert Merton first noticed that there exists an institutional norm in science that mandates widely sharing results. He called this the *communist norm*, according to which "[t]he substantive findings of science. . . are assigned to the community. . . The scientist's claim to 'his' intellectual 'property' is limited to that of recognition and esteem" (Merton 1942, p. 121). Subsequent empirical work by Louis et al. (2002) and Macfarlane and Cheng (2008) confirms that over ninety percent of scientists recognize this norm of sharing. Moreover, most scientists (if not as many as ninety percent) consistently conform to the communist norm.

The existence of this norm raises two questions. Where did it come from? And how does it persist? In light of what I said above, these are important questions. A good understanding of what makes the communist norm persist tells us which aspects of the institutional (incentive) structure of science can be changed without affecting the communist norm. Understanding its origins might allow us to reinstate the communist norm if it disappeared for whatever reason. Insofar as we value the existence and success of science, these are things we should want to know.

There must be some sense in which it is in scientists' interests to uphold the communist norm and conform to it, or else it would disappear. One such sense is given by Strevens (forthcoming). He gives what he calls a "Hobbesian vindication" of the communist norm by showing that scientists should be willing to sign a contract that enforces sharing. The claim is that, from a credit-maximizing perspective, it is not beneficial for an individual scientist to share her work (which would help other scientists more than her), but every scientist is better off if everyone shares than if no one shares.

As Strevens is well aware, this only partially answers the question of the persistence of the communist norm, and says little about its origins. In contrast, I argue that *sharing is rational from a credit-maximizing perspective for an individual scientist*. If my argument is successful, it provides a much more detailed account of both the origins and the persistence of the communist

norm. It also adds to a tradition of work in philosophy and economics that has emphasized how individual scientists' "selfish" desire to receive credit for their work furthers the aims of science (e.g., Kitcher 1990, Dasgupta and David 1994, Strevens 2003).

Because the existence of a norm can itself change what is in scientists' interests to do, the sense in which sharing is or is not rational or beneficial to scientists needs to be clarified. For this purpose, I rely on the terminology for social norms developed by Bicchieri (2006). I explain this terminology in section 2.2 and use it to state Strevens' position more precisely than I did above.

Section 2.3 sets out my own position by explaining how the idea that scientists can publish and claim credit for intermediate results can be used to establish the rationality of sharing. Section 2.4 makes this more precise by describing a game-theoretic model of scientists working on a research project needing to decide whether to share their intermediate results.[2]

I then show that rational credit-maximizing scientists should indeed be expected to share (section 2.5). In section 2.6 I use these results to give an explanation of the persistence of the communist norm, and I consider some objections. I extend my explanation to include the origins of the norm in section 2.7, which involves considering boundedly rational scientists and some historical evidence. A brief conclusion wraps up the paper.

## 2.2 Social Norms and Communism

The question that this paper focuses on is whether it is in a scientist's interest to behave in accordance with the communist norm. More specifically, would it be in scientists' interest to share their work even in the absence of a norm telling them to do so? To clarify the question, I use some terminology defined

[2]The idea of using game theory to get a better understanding of norms in science goes back at least as far as Bicchieri (1988).

by Bicchieri (2006). She defines a *social norm* as follows:

> Let $R$ be a *behavioral rule* for situations of type $S$, where $S$ can be represented as a mixed-motive game. We say that $R$ is a social norm in a population $P$ if there exists a sufficiently large subset $P_{\text{cf}} \subseteq P$ such that, for each individual $i \in P_{\text{cf}}$:
>
> *Contingency*: $i$ knows that a rule $R$ exists and applies to situations of type $S$;
>
> *Conditional preference*: $i$ prefers to conform to $R$ in situations of type $S$ on the condition that:
>
> (a) *Empirical expectations*: $i$ believes that a sufficiently large subset of $P$ conforms to $R$ in situations of type $S$;
>
> and either
>
> (b) *Normative expectations*: $i$ believes that a sufficiently large subset of $P$ expects $i$ to conform to $R$ in situations of type $S$;
>
> or
>
> (b′) *Normative expectations with sanctions*: $i$ believes that a sufficiently large subset of $P$ expects $i$ to conform to $R$ in situations of type $S$, prefers $i$ to conform, and may sanction behavior. (Bicchieri 2006, p. 11)

The crucial feature of this definition is the requirement of normative expectations. This says that an individual's preference to conform to the norm is conditional on others' expectations (possibly enforced by sanctions). For example, norms surrounding the sharing of food are plausibly social norms: in the absence of others expecting them to share, many people might prefer not to share even if they knew most other people shared. In contrast, if an individual knows that in a particular country most people drive on the right side of the road, she would probably prefer to conform to that even if others had no expectations about her behavior.

The language of game theory is useful to sharpen these ideas. Consider a situation of type $S$ and a behavioral rule $R$. Recall that conforming to $R$ constitutes a *(Nash) equilibrium* if no individual has an incentive to deviate unilaterally, i.e., everyone prefers to conform given that everyone else does.

If knowledge of $R$ and empirical expectations (that others will conform to $R$) are sufficient to make an individual prefer to conform to $R$, then according to this definition $R$ is an equilibrium of the underlying game that is being played in situations of type $S$. But if normative expectations are required, that is, if individuals only prefer to conform to $R$ if others expect them to conform (and, possibly, are willing to back this up with sanctions), then $R$ is not an equilibrium of the "original" game: it is only made into an equilibrium by the existence of the norm itself. So the existence of a social norm—unlike other kinds of norms, such as descriptive norms and conventions—transforms the underlying game by changing people's preferences, thus creating a new equilibrium (Bicchieri 2006, pp. 25–27).

Is the communist norm a social norm in this sense, i.e., are normative expectations a necessary ingredient to make it in scientists' interest to share their work? In order to answer this question, an account of scientists' interests is needed that is independent of the communist norm, so that the question can be asked whether a self-interested scientist would share her work in the absence of a normative expectation that she do so.

A scientist's achievements create for her a stock of *credit*. This credit is the means by which she advances her career, which determines both her income and her status in the profession. Insofar as a scientist is someone who is interested in building a career in science, it is then in her interest to maximize credit. This claim has been defended by philosophers and sociologists as diverse as Hull (1988, chapter 8), Kitcher (1990), Strevens (2003), Merton (1957, 1969), and Latour and Woolgar (1986, chapter 5).

This is not to deny that the scientist may have other interests, either as a scientist (e.g., to advance human knowledge) or apart from being a

scientist (e.g., to have time for other pursuits). But these are idiosyncratic. I aim to show that sharing is beneficial to scientists as a consequence of an interest that all scientists share. Credit maximization is, in my view, the only candidate here.

The institutions of science put a premium on originality. Credit is awarded to the first scientist to publish some particular result or discovery, and the amount of credit awarded is roughly proportional to the significance of the result. This feature of science is known as the *priority rule*, and the extent to which it shapes scientists' behavior is well-documented (Merton 1957, 1969, Kitcher 1990, Dasgupta and David 1994, Strevens 2003).

By rewarding only the first scientist, the priority rule encourages scientists to work and publish quickly (Dasgupta and David 1994). In this way, it seems that the priority rule creates an incentive for scientists to share their work. However, "the same considerations give you a powerful incentive not to share your results before you have extracted every last publication from them" (Strevens forthcoming, p. 2). If results were shared before publication, this would improve other scientists' chances of scooping important discoveries for which those results are relevant. So, Strevens argues, there is a split in the motivations provided by the priority rule:

> The priority rule motivates a scientist to keep all data, all technology of experimentation, all incipient hypothesizing secret before discovery, and then to publish, that is to share widely, anything and everything of social value as soon as possible after discovery (should a discovery actually be made). The interests of society and the scientist are therefore in complete alignment after discovery, but before discovery, they appear to be diametrically opposed. (Strevens forthcoming, pp. 2–3)

Of course, any sharing that happens after a discovery has been made does not help science in coming to that discovery faster. Thus, at the crucial

stage at which science can be sped up by sharing, the priority rule provides no incentive to do so, according to Strevens.

Strevens then goes on to show that a social contract, in which all scientists agree to widely share their work (even before discovery), would be beneficial to all scientists. In doing so, he shows that the problem of sharing has the structure of a Prisoner's Dilemma: every scientist would be better off if every scientist shared, but each individual scientist has an incentive not to share. The communist norm is thus a social norm on Strevens' view: without normative expectations to transform the game (into something that looks more like a Stag Hunt), widely sharing scientific work is not an equilibrium.

Strevens is not the only one to make this claim. For example, Resnik (2006, p. 135) observes that "the desire to protect priority, credit, and intellectual property" can motivate scientists to keep scientific results secret. Claims like this are also made by Dasgupta and David (1994, p. 500)[3], Arzberger et al. (2004, p. 146), Borgman (2012, p. 1072), and Soranno et al. (2015, p. 70), among others.

## 2.3 Communism and Intermediate Results

In this paper I argue that, given the priority rule, it is in a scientist's own interest to share her work widely, at least whenever she expects other scientists to do the same. In other words, sharing widely is an equilibrium of the relevant game even in the absence of normative expectations. The problem of sharing is thus not like a Prisoner's Dilemma: the role of the communist norm is not to change scientists' preferences to make sharing attractive (at least not primarily).[4] It merely describes a rule of behavior that it is in

[3]Dasgupta and David (1994, p. 502) go on to semi-formally characterize a situation very similar to the model of Boyer (2014) and this paper, but they draw the opposite conclusion: they agree with Strevens that the underlying game has the structure of a Prisoner's Dilemma.

[4]I do not deny that normative expectations calling on scientists to share their work exist, as they in fact appear to do (Louis et al. 2002, Macfarlane and Cheng 2008). The

scientists' own best interests to follow.

An important part of my argument is the insight that major discoveries can often be split into multiple smaller discoveries that were made along the way. Newton's famous comment "If I have seen further it is by standing on the shoulders of giants" illustrates this accumulative nature of science. Boyer (2014, p. 18 and p. 21) gives some more detailed examples: the construction of the first laser can be split into a theoretical development and the actual construction based on that theory, and the experimental test of the EPR thought experiment by Aspect et al. (1982) was preceded by a number of papers defining and refining the experiment.

In these cases each of the smaller discoveries was published as soon as it was done, rather than after the major discovery was completed. It is not obvious that the scientists involved were acting in their own best interest. On the one hand, credit can be claimed for the smaller discovery. On the other hand, the advantage that the smaller discovery gives on the way toward the major discovery is lost by publishing (and hence widely sharing) it. In fact, Schawlow and Townes seem to have lost the race to build the first working laser at least partially because their publication of the theoretical idea spurred on other teams.

Boyer (2014) provides a model to analyze this tradeoff. He shows that in some idealized circumstances the benefits of sharing these *intermediate results* outweigh the costs, with costs and benefits both measured in credit assigned via the priority rule. Although Boyer does not specifically discuss the communist norm, his result could be used to argue that normative expectations are not necessary to explain it: the priority rule encourages wide sharing of scientific work even before the potential of future discoveries based on this work has been exhausted, i.e., "before you have extracted every last publication" (Strevens forthcoming, p. 2).

---

point is rather that these are not required to explain the origins or persistence of the norm. I return to this point in section 2.6.

Unfortunately, things are not that simple. A number of objections can be made. The remainder of this section describes two such objections, which motivate the construction and analysis of a formal model in sections 2.4 and 2.5. In section 2.6 I flesh out the explanation of the communist norm suggested by this model, and I consider some further objections.

One may worry that Boyer's result is not general enough to support claims about the origins or persistence of the communist norm. By his own admission, he only shows that "there exist simple and plausible research situations for which the [credit] incentive to publish intermediate steps is sufficient" (Boyer 2014, p. 29). I aim to show that in fact all or most research situations are such that there is a credit incentive to publish intermediate results, which requires a more general model. The results I obtain may be viewed as generalizations of Boyer's—relaxing the assumptions that there are only two scientists, that the scientists are equally productive, and that scientists share either all or no intermediate results—although speaking strictly mathematically they are not (because Boyer uses discrete time steps and I use continuous time).

The second worry questions the relevance of equilibria. This worry has two sides. One side claims that showing that sharing is an equilibrium is not sufficient to show that one should expect real scientists to share, especially when there are also other equilibria (this is known as the equilibrium selection problem). The other side claims that observed behavioral patterns need not be the equilibrium of some underlying game. I alleviate both of these worries by showing that sharing is not merely an equilibrium, but an equilibrium that one should expect to be realized by both fully rational and boundedly rational scientists. Thus, the particular equilibrium considered here has behavioral implications.

## 2.4 A General Game-Theoretic Model of Intermediate Results

The game-theoretic model I develop in this section is intended to investigate scientists' incentives when they are working on a project that can be divided into a number of intermediate stages.[5] An *intermediate stage* is a part of the project which, when completed successfully, yields a publishable intermediate result in the sense of Boyer (2014, section 2). I assume that stages can only be completed in one order.[6] The number of intermediate stages of the project is denoted $k$.

Competition plays a central role in the model. I assume that scientists are aware that other scientists are working on the same project (or at least believe this to be the case). Merton (1961) argued for the ubiquity of multiple discoveries in science, which suggests that scientists should almost always expect other scientists to be working on the same project. I thus assume that $n \geq 2$, where $n$ is the number of scientists (or research groups) working on the project. Note that "scientist" may refer to someone working in the natural sciences, the social sciences, the humanities, or any other field where the priority rule applies.

Whenever a scientist completes an intermediate stage, she has to make a choice: she can either publish the result, or keep it to herself.[7] Publishing

[5] Although it was developed independently, the model turns out to be essentially identical to the model studied by Banerjee et al. (2014). In section 2.6 I briefly discuss their main result, which is roughly speaking a weaker result in a more general model. Banerjee et al. do not, however, give the detailed defense of the assumptions I give in this section, or the application to explaining the communist norm I give in sections 2.6 and 2.7.

[6] This linearity assumption may seem restrictive and unrealistic. However, any alternative assumption would only make sharing more attractive by improving the chance that a scientist can claim credit for an intermediate result without hurting her chances of future credit (because, e.g., other scientists are following a different path within the research project and thus are not helped by the publication of the intermediate result). Banerjee et al. (2014) consider this possibility in more detail.

[7] By assumption, the result is publishable, i.e., if she decides to publish it, it will be accepted by a journal.

benefits the scientist, because she thereby claims credit for completing that intermediate stage as well as any preceding stages that remain unpublished, in accordance with the priority rule. I assume that all stages are equally valuable, so the amount of credit obtained is equal to the number of stages published. Publishing also benefits the scientific community: other scientists no longer need to work independently on the stages that have been published. Publishing thus "expedites the flow of knowledge". I use $E$ to denote this strategy.

The scientific community's benefit is a potential downside to the individual scientist: if she keeps her results secret instead, she can start working on the next stage before anyone else can. This improves her chance of being the first to successfully complete the next stage, thus allowing her to claim credit for more stages later (at the risk that someone else claims credit for the one she did not publish). Holding onto a discovery until a more expedient time might thus be beneficial to the scientist. Call this strategy $H$.

When a scientist completes the last stage there is no incentive (within the model) to keep her from publishing. So when a scientist completes stage $k$ she always publishes, claiming credit for all unpublished stages and concluding this instance of the model.

Note that I assume that scientists care only about credit and that the only way to get credit is by publishing. Scientists are thus not assumed to have any inherent preference for or against sharing their work. In particular, expectations (normative or otherwise) from other scientists are not built into the individual scientist's preferences.

An interesting feature of the priority rule is its uncompromising nature. According to the priority rule, there are no second prizes, even if the time interval between the two discoveries is very small. This feature was noted by Merton (1957, p. 658), who quotes the French scientist François Arago as saying: "'about the same time' proves nothing; questions as to priority may depend on weeks, on days, on hours, on minutes."

To incorporate this feature in the model, it needs to be able to distinguish arbitrarily small time intervals. This suggests a continuous-time model: a model using discrete time units might place two discoveries in the same time unit even though in reality one of them happened (slightly) earlier than the other. This would fail to adequately model the uncompromising nature of the priority rule.

This means that a continuous-time probability distribution is needed to model the *waiting time*: the time it takes a given scientist to complete an intermediate stage. For this purpose I use the exponential distribution, the only candidate that has significant empirical support behind it (Huber 2001, more on this below). In particular, I assume that the time scientist $i$ takes to complete any intermediate stage follows an exponential distribution with parameter $\lambda_i$. The parameter can be interpreted as the average number of stages completed by the scientist per unit time. The parameter may be different for different scientists, indicating the possibility that some scientists work faster than others, or are part of a larger or more efficient research group.

I assume that the completion times for different scientists are probabilistically independent. In doing so, I set aside cases in which some event (external to the model) acts as a common cause that may lead multiple scientists to complete a stage around the same time (or prevent them from doing so).

The assumption that waiting times are exponential is equivalent to the assumption that scientists' productivity is a Poisson process with a parameter that is constant over time. Empirical work has shown that scientists' productivity fits a Poisson distribution quite well, and the percentage of authors who experience significant trends or surges over time is small. Huber (1998a,b) has established this for the rate at which patents are produced by inventors, Huber and Wagner-Döbler (2001a) for publications in mathematical logic, Huber and Wagner-Döbler (2001b) for publications in 19th century physics, and Huber (2001) for publications in modern physics, biology, and

psychology.[8]

The assumption that waiting times are exponential means that the probability that it will take scientist $i$ more than $t$ time units to complete a given stage is $\exp\{-t\lambda_i\}$.[9] This distribution has some formal features that I will make use of (Norris 1998, section 2.3). First, it is "memoryless". This means that after a certain amount of time has passed and the waiting time has not ended yet, the distribution of the remaining waiting time is equal to the original distribution of the waiting time. Second, the minimum of $n$ independent exponential random variables with parameters $\lambda_i$ $(i = 1, \ldots, n)$ is itself exponentially distributed with parameter $\lambda = \lambda_1 + \ldots + \lambda_n$. Thus, the waiting time until at least one of the scientists completes a stage of the project is exponentially distributed with parameter $\lambda$. Third, the probability that scientist $i$ is the first one to finish the stage she is working on is $\lambda_i/\lambda$.

The memorylessness property may seem odd, as it suggests that the scientist herself never knows whether she is making any progress on the problem. Moreover, if she starts working on a given stage much later than another scientist she has the same chance of completing it first as she would have had

[8]The fact that scientists' total career productivity follows a Poisson distribution (if accepted) does not imply exponential waiting times. One could generate Poisson distributions in other ways. But the evidence regarding trends and surges, as well as the fact that the evidence includes scientific careers cut short, suggests the stronger claim that at any given time in a scientist's career the Poisson distribution is a good model for her productivity up to that point. On this interpretation it is a simple mathematical consequence that the waiting times are exponential.

[9]Compare this with Boyer's assumption that there is a fixed probability $\lambda$ that a given scientist will solve a given stage in a given time unit. As noted above, by using discrete time units this model provides no way of applying the priority rule when two scientists finish the same stage in the same time unit. This problem can be addressed by using smaller time units. Suppose that what was previously one time unit is now $x$ time units, and in each unit the scientist completes the stage with probability $\lambda/x$. The probability that the scientist has not completed the stage at time $t$ (where $t$ is measured in the original time units before magnification) is $(1 - \lambda/x)^{tx}$. A continuous-time model is obtained by taking the limit as $x$ goes to infinity. Then the probability that the scientist has not completed the stage at time $t$ is $\lim_{x\to\infty}(1 - \lambda/x)^{tx} = \exp\{-t\lambda\}$. So, in addition to being independently empirically justified, exponential waiting times naturally arise as the limiting case of Boyer's model with continuous time.

if both had started at the same time (conditional on the fact that the other scientist does not complete the stage before she starts).

While these features of the exponential distribution do not seem to mesh well with the subjective experience of working on a research project, I want to insist that Huber's empirical evidence should be given more weight than subjective experience. The following consideration may help to smooth the apparent conflict.

In the model, scientists only make decisions after they have just completed a stage. So for the model it only matters that when a scientist completes a stage, she views the time she needs to complete future stages and the time other scientists need to complete stages as exponentially distributed. I do not need to insist that the scientist views the time needed to complete stages as exponentially distributed while she is in the middle of one.

How does my model compare to the one given by Strevens (forthcoming)? Perhaps the key difference is that contrary to Strevens I have described a zero-sum game. In my model it is implicitly assumed that the scientists will eventually complete the entire research project.[10] Since each stage is worth one unit of credit, and the first scientist to complete stage $k$ always claims credit for it and any unclaimed stages, this means that at the end of the game the scientists have always divided $k$ units of credit between them. So any change in strategy that leads to one scientist improving her (expected) credit must lead to a decrease for at least one other scientist.

In contrast, a key component of Strevens' model is the chance each scientist has of successfully completing the research project "in isolation", which leaves room for the scenario in which the research project is never completed by anyone. By sharing their progress, Strevens assumes, the scientists improve each other's chances of completing the research project. In fact this appears to be the main driving force behind his result that scientists should be willing to sign a social contract that enforces sharing: in his model sharing

[10]More precisely: the scientists complete all $k$ stages in finite time with probability one.

"creates" expected credit (by improving the overall chance that any credit is awarded at all), and as long as this "extra" credit is divided in such a way that everyone benefits at least a little (in expectation), it is clear that everyone will be better off if everyone shares.

By allowing for a chance that no scientist completes the research project, Strevens' model is arguably more realistic than mine. But I claim that this is a strength rather than a weakness of my model. Working with a zero-sum game reflects a strictly more pessimistic assumption about the benefits of sharing than working with a model like Strevens'. The result that sharing is incentive-compatible which I state and prove below is thus a somewhat surprising result: it is stronger than the result Strevens proved, while his model makes a more optimistic assumption about the benefits of sharing. Insofar as I show that the priority rule is sufficient for a communist norm to arise (without a need for normative expectations) in my model, this result should hold *a fortiori* in a more realistic (not zero-sum) model.

There are other ways to change the model that would make it no longer zero-sum. For example, Boyer-Kassem and Imbert (2015, section 4) argue that one should consider credit per unit time (rather than "total credit" which I use). Then sharing benefits all scientists to some extent by decreasing the expected completion time of the research project; Boyer-Kassem and Imbert call this a "speedup effect". So considering credit per unit time instead of total credit also invalidates the zero-sum property. Since, as above, it does so in a way that makes sharing more attractive, the result I get in my model holds *a fortiori* when credit per unit time is used.

## 2.5 The Incentive to Share in the Model

The previous section described a game-theoretic model of scientists working on a project that requires some number of intermediate stages to be completed. The game consists of a sequence of (probabilistic) events, in which

the scientists can intervene at specific points through their choice of strategy by publishing their work ($E$) or keeping it secret ($H$). Each scientist attempts to maximize her credit.

In the simplest version of the game there are two scientists ($n = 2$) and the research project has two stages ($k = 2$). The extensive form of the game is given in figure 2.1.

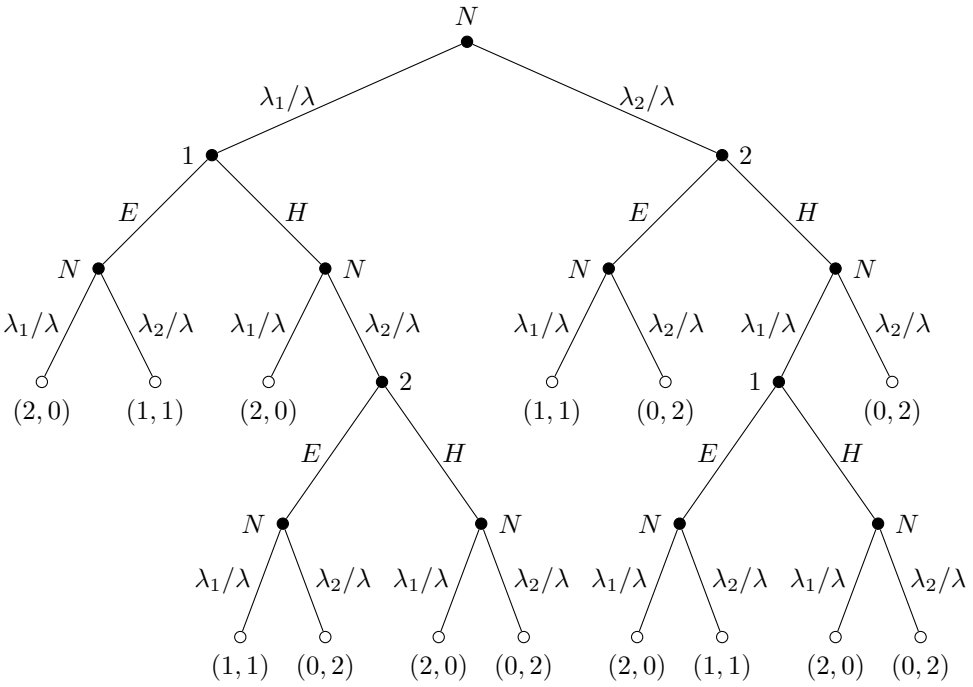


Figure 2.1: Extensive form of the game with $n = 2$ and $k = 2$

At the root node Nature decides which of the two scientists is the first one to complete the first stage of the project with the indicated probabilities. This leads to one of two decision nodes marked with a number indicating which scientist makes a decision at this node.

If the scientist publishes (strategy $E$), she collects one unit of credit. Both scientists now know the solution to stage 1 of the project, so they start working on stage 2. Nature decides which of the two scientists completes the

second stage first. Then the game ends, with payoff pairs indicating credit awarded to each scientist (one unit for each published stage).

What if a scientist chooses not to publish her solution to the first stage of the project (strategy $H$)? Then she does not collect a unit of credit, and the other scientist does not learn the solution to stage 1. So now one scientist starts working on stage 2, while the other continues working on stage 1. Nature decides which of the two scientists finishes the stage she is working on first (due to the memorylessness of the exponential distribution, the scientist working on stage 1 is not more likely to finish fast). If Nature picks the scientist working on stage 1 the game ends and that scientist gets both units of credit.

Otherwise, the other scientist get a chance to claim the first unit of credit by playing strategy $E$, or defer by playing $H$. In either case, both scientists can now work on stage 2. The first scientist to complete stage 2 receives either one unit of credit (if the solution to stage 1 has been published already) or both units of credit (if both scientists played strategy $H$).

It is implicitly assumed in figure 2.1 that scientist 1 knows when scientist 2 completes a stage, even when she keeps the result secret. Is it realistic to assume that scientists have this kind of information? I think this differs from field to field. In small fields where everyone knows what everyone else is working on word gets around when one of the labs has solved a particular problem, even when they manage to keep the details to themselves. Or, with pre-registration of clinical trials becoming more and more common, scientists might know that some other scientist knows, say, whether a particular drug is effective, without knowing whether the answer is yes or no.

But in other fields this kind of information might not be available. If this assumption is dropped scientists are unable to distinguish between certain decision nodes, indicated by so-called *information sets* (see figure 2.2). This yields a *game of imperfect information*.[11] In contrast, the version of the game

[11]This is a technical term for a game in which players cannot distinguish certain decision

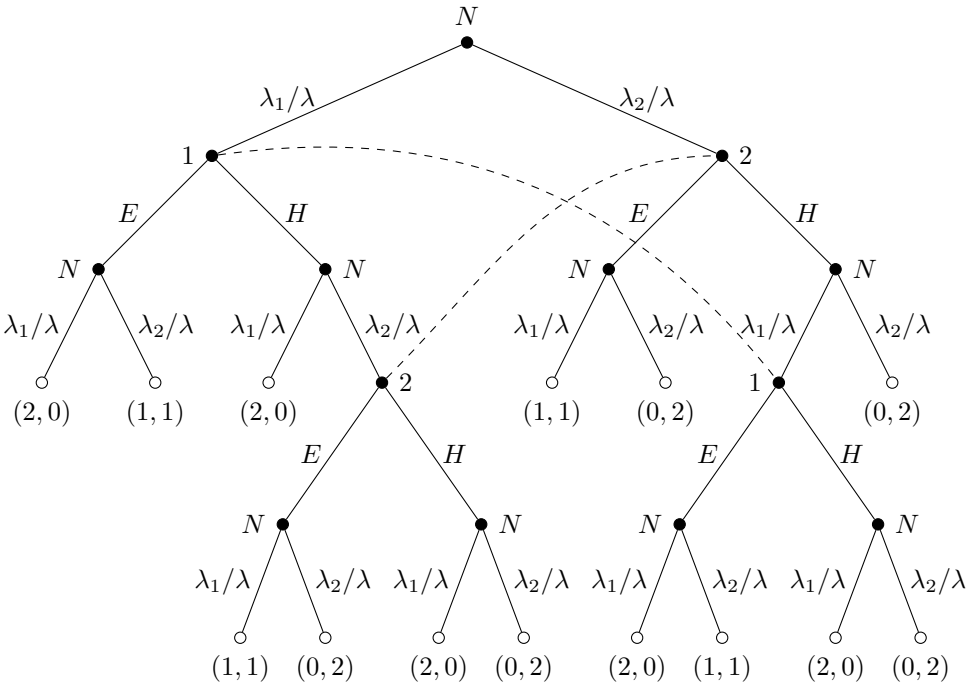


Figure 2.2: Extensive form of the game of imperfect information with $n = 2$ and $k = 2$

in which scientists can make these distinctions (as in figure 2.1) is a *game of perfect information*. I analyze both versions of the game.

Recall that I am interested in finding equilibria of these games. One way to find an equilibrium in a game of perfect information is by *backwards induction*. This involves identifying what a rational scientist will do at a terminal decision node, and then going backwards through the tree, identifying rational actions for the scientists by assuming other scientists will play rationally downstream.

In figure 2.1 it is rational for the scientists at the bottom nodes to play strategy $E$: this yields either the same payoff or a higher payoff than playing strategy $H$. Assuming that the scientists play $E$ at the bottom nodes, it is

nodes. Not to be confused with a game of incomplete information, where the players may not know each other's preferences or possible strategies.

also rational for the scientists at the top nodes to play strategy $E$. Thus, the backwards induction solution of this game is for both scientists to play $E$ at both of their decision nodes.

Nothing in this backwards induction analysis depended on the values of $\lambda_1$ and $\lambda_2$, or, as the following theorem shows, on the number of scientists or the number of stages. Moreover, any other equilibrium of the game is behaviorially indistinguishable from the backwards induction solution. That is, while there may be other equilibria, these differ only in that some scientists make different decisions at decision nodes that will not actually be reached in the game (see appendix A for a proof).

**Theorem 2.1.** *Consider the game with perfect information with $n \geq 2$ scientists and $k \geq 1$ stages.*

*(a) This game has a (unique) backwards induction solution in which all scientists play strategy $E$ at every decision node.*

*(b) There are no equilibria (in pure or mixed strategies) that are behaviorally distinct from the backwards induction solution.*

An equilibrium analysis thus yields a unique prediction for the game of perfect information. How about the game of imperfect information? Backwards induction does not apply to this type of game. But equilibria can still be identified using the expected payoff of each strategy. Table 2.1 gives the expected credit for each scientist when there are only two scientists and two stages (as in figure 2.2). Note that because the scientists cannot distinguish between their two decision nodes, only two (pure) strategies are available to them (compared to four in the game of perfect information).

This game has only one equilibrium regardless of the values of $\lambda_1$ and $\lambda_2$: both scientists play strategy $E$. Moreover, this is a *strict equilibrium*. An equilibrium is strict if, keeping the other scientists' strategies fixed, deviating from the equilibrium strictly decreases a scientist's expected credit. This is

| | $E$ | $H$ |
|---|---|---|
| $E$ | $\left(2\frac{\lambda_1}{\lambda}, 2\frac{\lambda_2}{\lambda}\right)$ | $\left(2\frac{\lambda_1}{\lambda} + \frac{\lambda_1^2}{\lambda^2}\frac{\lambda_2}{\lambda}, 2\frac{\lambda_2}{\lambda} - \frac{\lambda_1^2}{\lambda^2}\frac{\lambda_2}{\lambda}\right)$ |
| $H$ | $\left(2\frac{\lambda_1}{\lambda} - \frac{\lambda_1}{\lambda}\frac{\lambda_2^2}{\lambda^2}, 2\frac{\lambda_2}{\lambda} + \frac{\lambda_1}{\lambda}\frac{\lambda_2^2}{\lambda^2}\right)$ | $\left(2\frac{\lambda_1^2}{\lambda^2} + 4\frac{\lambda_1^2}{\lambda^2}\frac{\lambda_2}{\lambda}, 2\frac{\lambda_2^2}{\lambda^2} + 4\frac{\lambda_1}{\lambda}\frac{\lambda_2^2}{\lambda^2}\right)$ |

Table 2.1: Expected credit for each scientist as a function of scientist 1's strategy (row) and scientist 2's strategy (column)

a stronger requirement than that for an equilibrium because that definition allows for cases in which a deviating scientist is equally well off. It turns out that both of these features generalize for different numbers of scientists and stages (see appendix A for a proof).

**Theorem 2.2.** *Consider the game with imperfect information with $n \geq 2$ scientists and $k \geq 1$ stages.*

*(a) This game has an equilibrium in which all scientists play strategy $E$ at every information set.*[12]

*(b) There are no other equilibria (in pure or mixed strategies).*

*(c) The equilibrium is strict.*

What do theorems 2.1 and 2.2 say about what it is rational for a scientist to do? They say that if not every scientist immediately shares any stage that she completes, there is at least one scientist who is irrational in the sense that she would have had a higher expected credit if she had played a different strategy. So the only way these scientists can all be rational is if they all share every stage. In other words, if all scientists are rational expected credit maximizers they will all share.

[12]This result is a corollary of Banerjee et al. (2014, theorem 2.1). See the discussion in section 2.6.

## 2.6 Explaining the Persistence of the Communist Norm

I take the results from the game-theoretic model obtained above to give an explanation for the *persistence* of the communist norm. The explanation runs as follows.

Suppose the communist norm is in place, i.e., scientists are sharing their intermediate results. If a given scientist deviates by not sharing an intermediate result, she thereby lowers her expected credit (this is just what it means for sharing to be a strict equilibrium). Hence the scientist has a credit incentive to return to conforming to the norm. So credit incentives can correct small deviations from the norm.

Note that I do not claim that real scientists are rational credit-maximizers. This is not necessary for my explanation. I have shown that rational credit-maximizing scientists would conform to the norm. All that follows for real scientists is that they have a credit incentive to conform to the norm. This fact, combined with the fact that real scientists are at least somewhat sensitive to credit incentives (more on this in section 2.7), constitutes my explanation of the persistence of the norm.

Here I want to point out a number of peculiar features of my explanation and consider some objections based on those features.

Because my explanation depends on the claim that it is rational for credit-maximizing scientists to share their intermediate results, which is supported by a game-theoretic model, the explanation's force depends on the generality of that model, and hence on the assumptions I made along the way.

A number of these assumptions are, perhaps surprisingly, not restrictive. These are: (1) the assumption that each intermediate stage is equally difficult, (2) the assumption that the relative productivity of scientists is the same for all stages, (3) the assumption that the reward for each stage is equal (one unit of credit), and (4) the assumption that the stages are "linear" (i.e.,

there is only one sequence of stages that leads to completion of the project).

Consider a version of the game with imperfect information in which scientists' productivity, as well as credit rewards, may be different for each stage. Write $\lambda_{i,j} > 0$ for scientist $i$'s productivity at stage $j$ and $c_j > 0$ for the reward for stage $j$. Let $\lambda_{-i,j} = \sum_{i' \neq i} \lambda_{i',j}$.

**Theorem 2.3** (Banerjee et al. (2014))**.** *There is an equilibrium in which all scientists play strategy E at every information set if the following condition holds: for any scientist i and for any pair of stages j and j′ such that j precedes j′,*

$$\frac{c_j \lambda_{-i,j}}{c_{j'} \lambda_{-i,j'}} \geq \frac{\lambda_{i,j'}}{\lambda_{i,j'} + \lambda_{-i,j'}}.$$

This result generalizes theorem 2.2.a: sharing remains an equilibrium as long as credit is given proportionally to the difficulty of the stage or earlier stages are rewarded relatively highly. Banerjee et al. (2014, theorem 3.2) provide a further generalization which allows arbitrary (acyclic) networks of stages.

These results come with a caveat. Because Banerjee et al. (2014) show neither uniqueness[13] nor strictness, persistence of the communist norm is only guaranteed when deviations from the norm are small, and the argument I give below for the origins of the communist norm does not work at all. Hence I focus in this paper on the more restricted model.

Neither my results nor theorem 2.3 cover the case in which later stages have a higher credit value. But it seems quite realistic that the scientist to finish the last stage ("puts it all together") might get more credit. These are exactly the circumstances in which there may be an incentive not to share.[14]

[13] Banerjee et al. only prove uniqueness for a case in which some of the scientists commit to sharing before the game starts (so-called Stackelberg agents). Theorem 2.2 above does not require this.

[14] Boyer (2014, theorem 3) suggests (for the case where $n = 2$ and $k = 2$) that if the second stage is worth up to twice as much credit as the first, there may still be an incentive to share. This would indicate some fairly significant robustness of the result.

From a descriptive perspective, these might be the kinds of cases where scientists do not share their intermediate results, and with good reason. From a normative perspective, this could be viewed as an argument against giving extra credit to the scientist who finishes the last stage (because the more equal the division of credit, the more incentive scientists have to share, as Boyer's, Banerjee et al.'s, and my results all suggest).

My explanation relies on three basic principles: scientists' sensitivity to credit incentives, the credit-worthiness of intermediate results, and the priority rule as the mechanism for assigning credit. These ingredients are sufficient to explain the persistence of the norm. In particular, there is no need for a social contract, normative expectations, or altruism.

This leads to a potential objection. On my construal, the communist norm is not strictly a social norm in Bicchieri's sense, as normative expectations have no role in the explanation. But the available evidence seems to refute this: scientists do view the communist norm as a social norm, they (normatively) expect other scientists to conform to it, and they feel the weight of this expectation when making their own decisions (Louis et al. 2002, Macfarlane and Cheng 2008). This appears to be at odds with my model: since the game is zero-sum, other scientists actually benefit when a given scientist deviates from the norm, so from a credit-maximizing perspective they should be encouraging each other to deviate.

But the model considers only those scientists who are directly competing on a given research project. While those scientists may stand to gain if their competitors fail to share their intermediate results, the wider scientific community stands to lose, as it will take longer to complete the research project. It is this wider community, I claim, that is the source of any normative expectations regarding sharing behavior. The normative expectations can then also be explained from self-interest, as the completion of the research project

But this turns out to be an artifact of Boyer's assumption that the scientists have equal productivity. Depending on the differences in productivity, the robustness of the result may be arbitrarily small (Banerjee et al. 2014, corollary 2.2).

may benefit other scientist' research.[15]

This yields an empirical prediction that might be used to help decide between Strevens' explanation and mine. On Strevens' explanation a deviation from the communist norm is a breach of a social contract which most directly impacts the immediate competitors of the scientist within the research project, who may legitimately regard it as unfair. On my explanation a deviation actually benefits the immediate competitors; the most direct negative impact is on those scientists who work on nearby projects. An examination of which scientists (direct competitors or those working on nearby projects) tend to most vocally object to deviations from the communist norm may thus shed light on the question which of these explanations is closer to the truth.

Another feature of my explanation is that it explains sharing behavior only for "intermediate results", i.e., results that are significant enough to be publishable in their own right. Strevens points out that on this view, "nothing will be shared until something relevant is ready for publication, and worse, it is only what characteristically goes into the journals that gets broadcast, so details of experimental or computational methods and raw data will remain hidden" (Strevens forthcoming, p. 5). This constitutes an objection to my explanation, as according to Strevens the communist norm requires that any and all results should be shared, regardless of their credit-worthiness.

To this worry I reply that it is not clear that the communist norm makes such strong requirements. When the material under consideration is too little or too detailed to be considered publishable, scientists' actual compliance with a putative norm of sharing drops off steeply (Louis et al. 2002, Tenopir et al. 2011).[16] If Strevens' aim is to explain a norm of sharing for these cases,

[15] Alternatively or additionally, normative expectations may arise simply because everyone in the community is in fact behaving in a certain way. Bicchieri points out that "[s]ome conventions may not involve externalities, at least initially, but they may become so well entrenched that people start attaching value to them" (Bicchieri 2006, p. 40).

[16] If it is assumed that material that cannot be published in a journal is worth zero credit when shared, then my model would predict that nothing would be shared. This

he may be trying to explain something that does not exist.

Strevens may reply to this that regardless of the content of the norm currently in place, it would be good to have a maximally inclusive communist norm. After all, scientists would benefit most from each other's work (thus speeding up the overall progress of science) if they shared results even before they had achieved publishable size and without hiding crucial details. By using the framework of a social contract to point out the benefits of more widespread sharing, Strevens could argue, it might be possible to help the scientific community get to such an improved norm.

That would be a laudable goal. However, the results from my model can do the same. They suggest a clear way to make it incentive-compatible for scientists to share work below publishable size: allow smaller publications. And sharing crucial details can similarly be made incentive-compatible just by giving credit for it (Tenopir et al. 2011, Goring et al. 2014). If getting scientists to share these minor results or crucial details is a goal that scientists and policy makers consider important, the model gives clear directions on how to get there (but it may not be possible or desirable to do this, cf. Boyer 2014, section 4.4). Modern information technology readily suggests ways in which this can be done without overburdening existing scientific journals. Developments in this area are already underway (Piwowar 2013). In this sense, the results from this paper are more actionable than Strevens'.

## 2.7 Explaining the Origins of the Communist Norm

Above I argued that the results from the game-theoretic model explain the persistence of the communist norm. It could be argued that they also explain

prediction is not borne out empirically: while there is much less sharing of this kind of material, there is still some sharing. Perhaps this behavior is simply unexplainable from a pure credit-maximizing perspective. However, the assumption that this material is worth zero credit may not need to be granted. See Piwowar (2013) and the discussion below.

the *origins* of the norm: the uniqueness clauses in theorems 2.1 and 2.2 guarantee that behavior in accordance with the communist norm is the only pattern that rational credit-maximizing scientists could settle on.

But such an argument would make stringent demands on the scientists' rationality which real scientists are unlikely to satisfy. This section investigates the question whether less than perfectly rational scientists would also learn to share their intermediate results, thus giving a more robust account of the origins of the communist norm.

To answer this question I consider a boundedly rational learning rule that makes only minimal assumptions on the cognitive abilities of the scientists. In particular, it requires only that the scientists know which strategies are available to them and that they can compare the credit earned on the previous round to that earned on the current round (where a "round" is one instance of the game of imperfect information; to evaluate this bounded rationality rule one needs to assume the game is played repeatedly).

The rule I consider is *probe and adjust*. A scientist using probe and adjust follows a simple procedure: on each round, play the same strategy as the round before with probability $1 - \varepsilon$, or "probe" a new strategy with probability $\varepsilon$ (with $0 < \varepsilon < 1$; $\varepsilon$ is usually "small"). In case of a probe, she picks a new strategy uniformly at random from all possible strategies. After playing this strategy for one round, the probe is evaluated: if the payoff for the round in which she probed is higher than the payoff in the previous round, keep the probed strategy (at least until the next probe); if the payoff is lower, return to the old strategy; if payoffs are equal, return to the old strategy with probability $q$ and retain the probe with probability $1 - q$ $(0 < q < 1)$.[17]

Consider a population of $n \geq 2$ scientists using probe and adjust to de-

[17]Note that this is not quite the same as asking whether the probed strategy is a better reply to the other scientists' strategy than the old strategy, as other scientists may have changed their strategy as well. In particular, if all scientists are using probe and adjust, simultaneous probes and probes on subsequent rounds prevent this rule from necessarily always picking the better reply.

termine their strategy in repeated plays of the game of imperfect information with the number of stages $k \geq 1$ fixed. Assume all scientists use the same values of $\varepsilon$ and $q$ (this assumption can be relaxed, see Huttegger et al. 2014, pp. 837–838). Then the following result can be proven (see appendix A).

**Theorem 2.4.** *For any probability $p < 1$, if the probe probability $\varepsilon > 0$ is small enough there exists a $T$ such that on an arbitrary round $t$ with $t > T$, all scientists play strategy $E$ at every information set with probability at least $p$.*

If, on a given round, all scientists play strategy $E$ at every information set, they may be said to have learned to share their intermediate results. The theorem says that the probability of this happening can be made arbitrarily high by choosing a small enough probe probability and a long enough waiting time. Moreover, the theorem says that once the scientists learn to share their intermediate results they continue to do so on most subsequent rounds. So even on this cognitively simple learning rule both the origins and the persistence of the communist norm can be explained on the basis of credit incentives.

Having already shown the same to be the case for highly rational scientists in section 2.5, I suggest that similar results should be expected for intermediate levels of rationality.[18] Conforming to the communist norm is then shown to be incentive-compatible for credit-maximizing scientists regardless of their level of rationality.

How historically plausible is my claim that credit incentives are responsible for the origins of the communist norm? It is not entirely clear how one should evaluate this question. But a necessary condition for my explanation to be correct is that credit for scientific work, and in particular credit awarded in accordance with the priority rule, predates the communist norm. The remainder of this section argues that this condition is satisfied.

[18] Because the equilibrium in the game of imperfect information is both strict and unique, various other learning rules and evolutionary dynamics can easily be shown to converge to it. Examples include fictitious play, the best-response dynamics, and the replicator dynamics.

As Merton (1957) points out, scientists' concern for priority goes back at least as far as Galileo. In 1610, he used an anagram to report seeing Saturn as a "triple star" (the first sighting of the rings of Saturn). The device of the anagram served "the double purpose of establishing priority of conception and of yet not putting rivals on to one's original ideas, until they had been further worked out" (Merton 1957, p. 654). If Galileo was concerned about establishing priority for his ideas, it seems that the priority rule must already have been in effect in 1610. Priority disputes also go back at least as far, as Galileo wrote multiple polemics to defend his priority on various discoveries (Merton 1957, p. 635).

The communist norm, on the other hand, was not established as a norm of science until around 1665. At the time, "many men of science still set a premium upon secrecy" (Zuckerman and Merton 1971, p. 69). The first scientific journals—the *Journal des Sçavans* and the *Philosophical Transactions*, both founded in 1665—were instrumental "for the emergence of that component of the ethos of science which has been described as 'communism': the norm which prescribes the open communication of findings to other scientists" (Zuckerman and Merton 1971, p. 69).

## 2.8 Conclusion

In the introduction I argued that the sharing of scientific results (mandated by the communist norm) is important to the success of science and indeed to the existence of science as we know it. My results show that the priority rule gives scientists an incentive to share any and all intermediate results. These results can be used to explain both the origins and the persistence of the communist norm, answering the questions I raised in the introduction.

If my explanation is accepted, the crucial features of the social structure of science that maintain the communist norm are seen to be the fact that scientists respond to credit incentives, the priority rule, and the credit-

worthiness of intermediate results. Tinkering with these features thus risks undercutting one of the most central aspects of science as a social enterprise.

By emphasizing credit incentives moderated by the priority rule, this paper falls in the tradition of Kitcher (1990), Dasgupta and David (1994), and Strevens (2003). Like those papers, I have picked one aspect of the social structure of science, and shown how the priority rule has the power to shape that aspect to science's benefit.

I take my results to show that no special explanation (using, e.g., normative expectations and/or a social contract) is required for the communist norm, *contra* Strevens (forthcoming). However, this only applies to whatever is publishable (or otherwise credit-worthy) in a given scientific community. Sharing scientific work that is too insignificant to be published is not incentivized in the same way. But insofar as this is a problem it suggests its own solution: give credit in accordance with the priority rule for whatever one would like to see shared, and scientists will indeed start sharing it.

# Chapter 3

# Expediting the Flow of Knowledge Versus Rushing into Print

## 3.1 Introduction

The desire to be first and receive *credit* for their work are important motivations for scientists. This is not to deny that scientists may have other motivations, such as advancing the state of human knowledge. But such motivations are idiosyncratic, while credit is necessary to have a career in science, and so professional scientists have to care about it to some extent. This point has long been recognized by sociologists such as Merton (1957, 1969) and Latour and Woolgar (1986, chapter 5) and philosophers of science like Hull (1988, chapter 8), Kitcher (1993, chapter 8), and Strevens (2003).

Recognizing scientists' desire for credit may seem problematic on a naive picture of science in which scientists selflessly strive for truth (see Kitcher 1993, chapter 1, for a caricature with references). A trend in the *social epistemology of science* has been to argue that things are not so bad. Kitcher (1993, chapter 8) and Strevens (2003) argue that credit can incentivize scien-

tists to distribute themselves over research programs in a way that is closer to optimal than if they were individually epistemically rational. Dasgupta and David (1994) argue that credit incentives speed up the progress of science. Boyer-Kassem and Imbert (2015) argue that credit incentives encourage collaboration between scientists. And Boyer (2014), Strevens (forthcoming), and chapter 2 of this dissertation argue that credit incentives can motivate scientists to share their work widely.

Given this trend, it would be tempting to conclude that credit incentives act like an infallible "invisible hand" that can solve all problems in the social organization of science. Here, I go against this trend in identifying one way in which credit incentives may lead to bad outcomes at the social level.[1] The central claim of this paper is that credit incentives create a pressure to publish that, when combined with a system of peer review that is necessarily imperfect, leads to reproducibility problems.

The issue of *reproducibility* has come under increased scrutiny (Ioannidis 2005, Pashler and Wagenmakers 2012). Recent studies have shown that published work in fields such as medicine and psychology frequently fails to be reproducible (Prinz et al. 2011, Begley and Ellis 2012, Open Science Collaboration 2015). Some have argued that the pressure to publish is a cause of this phenomenon (Fanelli 2010, Prinz et al. 2011). I illustrate this claim in section 3.2 with a case study in which the pressure to publish led to the publication of research that failed to reproduce: Fleischmann and Pons' work on cold nuclear fusion.

I then go on to argue that reproducibility problems arise as a structural problem of credit incentives rather than as (a series of) incidents. In section 3.3 I construct a model to show that two crucial ingredients are sufficient for reproducibility problems to arise. First, the system of peer review which aims to publish accurate results while rejecting erroneous results. Second,

[1] I am not the first to do so. For example, Strevens (2013) discusses ways in which credit incentives may lead to herding behavior, and Bright (forthcoming) argues that credit incentives are a cause of fraud in science.

the tradeoff between speed and reliability: the scientist must choose how quickly to work, knowing that the faster she works the greater the chance of errors. The following quote from a recent blog post illustrates how credit incentives may affect this tradeoff:

> [M]ost scientists have a little lab-coat scientist on one shoulder whispering in one ear ("Maybe you should take longer and replicate that result. Technically, you just inflated your alpha! Double-check those analyses. Maybe you should just re-run this experiment.") and a little tweed-jacket academic on the other shoulder whispering in the other ear ("Publish this before you get scooped. I already know the best spin. You need this on your CV for the next grant application.") (Carter 2015)

I show that the scientist has a unique credit-maximizing way to trade off speed against reliability. But, in a way to be made precise, the tradeoff favors speed over reliability. Scientists are given too much of an incentive to "rush into print", compared to what would be optimal from a social perspective. This shows that credit incentives are a possible cause of reproducibility problems.

In section 3.4 I expand the model by allowing the scientist to also choose a desired level of impact. High-impact work has greater scientific value and yields more credit, but this trades off against speed and/or reliability. I show the robustness of my earlier results in this expanded model, and I consider how it gives rise to different types of scientists: "impact-seekers" and "safety-seekers".

In the conclusion (section 3.5) I summarize my results. I also discuss possible ways to diminish or remove the incentive to produce work that is structurally less reproducible than is socially desirable. And finally, I discuss whether my results support an interpretation of Fleischmann and Pons' cold fusion research as a case of "rushing into print".

## 3.2 Cold Fusion

In this section I use a case study to argue that the pressure to publish can lead to the publication of research that fails to reproduce. The next section aims to show that this is a structural rather than an incidental problem.

On March 23, 1989, two established and respected chemists named Martin Fleischmann and Stanley Pons gave a remarkable press conference at the University of Utah (UU). They claimed that by loading a palladium rod with deuterium through electrolysis, they had turned the rod into a source of energy, producing up to four times as much heat as they put in.

They hypothesized that the deuterium atoms might be packed together so closely within the palladium as to force pairs of them together in an energy-producing process known as *nuclear fusion*. Conventional wisdom held that a sustained, controlled fusion reaction—the kind needed for a viable source of energy—requires temperatures over a hundred million degrees (among other things). Now two chemists claimed to be able to achieve the same thing at room temperature. Hence the phenomenon came to be known as *cold fusion*.

Although it seemed impossible given existing theories, physicists and chemists alike were initially willing to give Fleischmann and Pons the benefit of the doubt. Experimental results take precedence over theory, and the two's credentials as experimentalists were impeccable. Given the potential implications, and the media hype, scientists around the world dropped what they were doing to attempt to replicate the experiment.

Within the first few weeks after the press conference, a number of announcements were made (usually also via press conference) by researchers seeing similar phenomena. But as time passed their claims came under heavy criticism. The excess heat measurements were attributed to mistakes in accounting for the potential recombination of gases released during the experiment. The neutron measurements (Fleischmann and Pons' other important piece of evidence) could not be replicated with more sophisticated equipment. After the meeting of the American Physical Society (APS) in

May 1989, the tide shifted from a mixture of excitement and skepticism to a consensus that Fleischmann and Pons had been mistaken.

The current scientific consensus, then, is that it is not possible to achieve cold fusion at meaningful rates. Fleischmann and Pons' claim to the contrary on March 23, 1989, has been heavily criticized by scientists. In their books on the case, Close and Huizenga judge that they "went public too soon with immature results" (Close 1991, p. 328) and that their "gamble to go public... is the scientific fiasco of the century" (Huizenga 1993, p. 214). What led Fleischmann and Pons to make this fateful decision to "go public"?

While in the press conference they claimed to have been working on this project for five years, in reality most of the work had been done in the last six months (Close 1991, p. 82). Before then, they had done some exploratory work that seemed promising. In August 1988, they requested funding for their cold fusion research from the Department of Energy (Huizenga 1993, p. 16). This brought them into contact with Steven Jones, a professor of physics at Brigham Young University (BYU).

Unbeknownst to Fleischmann and Pons, Jones and his team had been working on a very similar project. The main differences were that Jones was primarily interested in explaining some of the heat at the center of the Earth (rather than creating a new source of energy) and that he focused on measuring neutron production rather than excess heat.

Jones noted that their work seemed complementary. The two teams exchanged information and Fleischmann and Pons visited Jones' lab on February 23, 1989. By now Jones had obtained his best data, which gave some evidence of neutrons above background levels at the right energy level to be potentially due to fusion. Jones announced that he was going to present his data at the APS meeting in May and was planning to submit an article to a journal soon.

Fleischmann and Pons, in contrast, were not ready to go public. Pons' graduate student Marvin Hawkins had been running experiments since Octo-

ber 1988. They were confident in their evidence that some experiments produced excess heat, but they had only just started measuring neutrons, and their apparatus was much less sophisticated than Jones'. Much remained to be investigated: why did some experiments produce excess heat while others did not, and how could they explain the discrepancy between the heat measurements and the neutron counts (which were many orders of magnitude too low compared to the heat if fusion was occurring)? Fleischmann and Pons indicated that they wanted to do another eighteen months of research before going public (Pool 1989, Huizenga 1993, p. 18).

With Jones about to go public, Fleischmann and Pons felt that they had to inform the university. This led to another meeting at BYU on March 6, this time with the presidents of both universities present. Especially at UU there was a clear sense of the potential impact of cold fusion at this point, with the president of UU suggesting that billions of dollars and Nobel prizes were at stake (Close 1991, p. 93). Moreover, those associated with the UU group accused Jones of stealing their ideas on at least two occasions in mid-February and mid-March (Close 1991, chapter 6). After Jones insisted that he was ready to publish his results, the two groups agreed to submit their results to *Nature* simultaneously on March 24.

Jones has claimed that there was a further agreement not to publicize the work until that time, but Pons has denied this (Pool 1989, Close 1991, p. 94). Either way, Fleischmann and Pons did publicize their work: they sent a hastily written manuscript to the *Journal of Electroanalytical Chemistry* on March 11, and they held the above-mentioned press conference, a day before the scheduled simultaneous submission.

Why was this press conference held? Ostensibly it was to correct "rumors, leaks, questions, and false information" that were already doing the rounds (Pool 1989, Huizenga 1993, p. 19). But it seems clear that the real reason was to establish priority for the cold fusion research, especially relative to Jones

(Huizenga 1993, p. 19).[2] This may seem unnecessary, as Jones' experimental results were quite different (measuring neutrons rather than heat) and of such a different order of magnitude as to hold no promise for a viable source of energy. Jones' publication would thus not appear to be a threat to the originality or importance of Fleischmann and Pons' work. But this was not so clear at the time, as Fleischmann explained later.

> [I]n the situation we were then in, we were obliged to tell the university of the work we had done and they perceived that they were obliged to go for patent protection at that time. We could not tell whether Jones had heat data or was planning to look for this. How could one tell? He was certainly thinking about fusion as a source of heat in the Earth. If he was going to say *that* in the paper, which was surely his intention to do, it would almost certainly destroy any possibility of patent protection (quoted in Close 1991, pp. 99–100).

Thus, both the decision to agree to publish simultaneously with Jones and the later decisions to submit to a different journal before Jones and hold a press conference were made out of a concern for priority. Fleischmann and Pons were aware that their results were still preliminary (they wanted to do another eighteen months of research before publishing anything) but went public anyway to establish priority, under pressure from university officials.

So a concern for priority led to the publication of research that turned out to be erroneous. While this is a particularly high-profile case, it has recently been suggested that erroneous research is quite prevalent in the scientific literature (Prinz et al. 2011, Begley and Ellis 2012, Open Science Collaboration

[2]The claim by UU officials that the reason for the press conference was "leaks" does not hold up. The evidence for the leak was an article on cold fusion in the *Financial Times* on the morning of the press conference. This was neither a real leak (Fleischmann and Pons had given permission for it), nor could it have caused the press conference to be held, as this had been scheduled two days earlier (Close 1991, pp. 101–102).

2015). The model of the next two sections aims to establish that this is a structural problem stemming from scientists' (credit) incentives. In doing so, the model also lends some support to the claim that Fleischmann and Pons did nothing irrational by going public when they did, despite Close and Huizenga's criticism of this decision (as I argue in more detail in section 3.5).

## 3.3 A Tradeoff Between Speed and Reliability

Here I develop a decision-theoretic model to evaluate decisions to go public with results of scientific research. By giving a model, I aim to show that the problem of erroneous results arises structurally rather than incidentally. This section considers only the tradeoff between speed and reliability, while section 3.4 also allows the potential importance or impact of the result to influence the decision.

Consider a scientist—or a team of scientists, such as Fleischmann and Pons—working on a research project. Why would she decide to publicize her work, say in the form of a journal article? As the case of Fleischmann and Pons illustrates, an important reason is to establish priority for the work.

As I mentioned in the introduction, scientists' concern for priority is well-documented and understandable, given its importance to scientific careers (Merton 1957, 1969). But even a scientist who only cares about the progress of science altruistically would be concerned about priority. After all, the primary way to contribute to the progress of science is to do original work that causes other scientists to learn something earlier than they otherwise would have (cf. Strevens 2003, p. 55). Thus, while I use the notion of credit to represent the concern for *speed* in the model, I claim that this model can still capture the motivations of scientists not primarily concerned about credit.

For these reasons, the scientist prefers to work and publish faster rather than slower (all else being equal). This is represented in the model by as-

suming that the scientist aims to maximize the amount of credit she accrues per unit time. E.g., if the amount of credit per publication is some fixed number $c$ (this assumption is relaxed in section 3.4), working and publishing twice as fast will double credit per unit time, all else being equal.

But of course all else is not equal. Working faster reduces *reliability*. By reliability I mean the amount of certainty with which the result of the research project (e.g., "cold fusion is a viable source of energy") is established. Loosely speaking I have in mind the probability that the result is true, given the amount of evidence the scientist has gathered at the time of publication.

But this will not do in a credit-maximization model, since credit is conferred socially by other scientists, and thus cannot be determined directly by (objective) truth.[3] So to be more precise: I call a scientific result *accurate* if it holds up as "true" (i.e., is not discredited) in the relevant scientific community in the mid-term, say for ten years after publication. Conversely, I call a result inaccurate or *erroneous* if it does not hold up in the community in the mid-term.[4] The *reliability* is then the scientist's subjective probability, given the evidence gathered at the time of publication, that the result is accurate.

In the model, the scientist chooses the desired reliability $p \in [0, 1]$. That is, the scientist works on her research project until she obtains a result that she thinks has at least probability $p$ of holding up as "true" in the community, at which time she publishes.

The reliability $p$ is given as a single number, but a number of considerations are folded together here. The scientist's estimate of how likely her result is to hold up in the community depends not only on the evidence she has collected, but also on who she expects to be working on the problem,

[3] I could get around this problem by simply defining to be true that which a given scientific community accepts at a given time, as some sociologists of science have done (e.g., Collins 1981, Latour and Woolgar 1986, Bloor 1991). I do not use this way of speaking to avoid the impression that I am committed to this as a substantive theory of truth.

[4] I consider cases in which research fails to be reproducible, as discussed in the introduction, to be a special case of this.

how likely these other scientists are to overturn her results (both of which may vary as a function of how much effort the scientist puts in right now). As a consequence the way the reliability is determined may be different in different research contexts.

Moreover, I assume that the scientific community delivers a unanimous and dichotomous judgment in the mid-term (either the result holds up or it does not) whereas in reality the community may be divided and individuals in the community may be uncertain in their judgment of the result.

Finally, the fact that the reliability is used as a probability raises a number of questions. How can scientists coherently put probabilities on mathematical hypotheses (given the usual assumption of logical omniscience)? What happens if new hypotheses are introduced or other changes to the probability space occur?

Giving full consideration to all these factors requires a game-theoretic model in which the judgments and actions of the members of the scientist's community are fully represented. Here I make the simplifying assumption that in individual cases, the scientist can competently assess the reliability of her work, and a single number that works like a probability provides a reasonable approximation of that.

Reliability takes time. This is reflected in the model by the *speed function* $\lambda$. The value $\lambda(p)$ reflects the speed at which the scientist works if the desired reliability is $p$. More specifically, $1/\lambda(p)$ is the expected time until completion of the research project (so $\lambda(p)$ is the number of research projects "like this one" that the scientist would expect to complete per unit time). If $\lambda$ is a decreasing function of $p$, as I assume, the expected time until completion indeed goes up if $p$ is increased (see figure 3.1).

On the other hand, reducing the reliability (lowering $p$) allows the scientist to work more quickly. This represents the idea of "rushing into print", leading to a higher probability of errors slipping in. I assume that these errors represent honest mistakes by the scientist. The present model is not intended

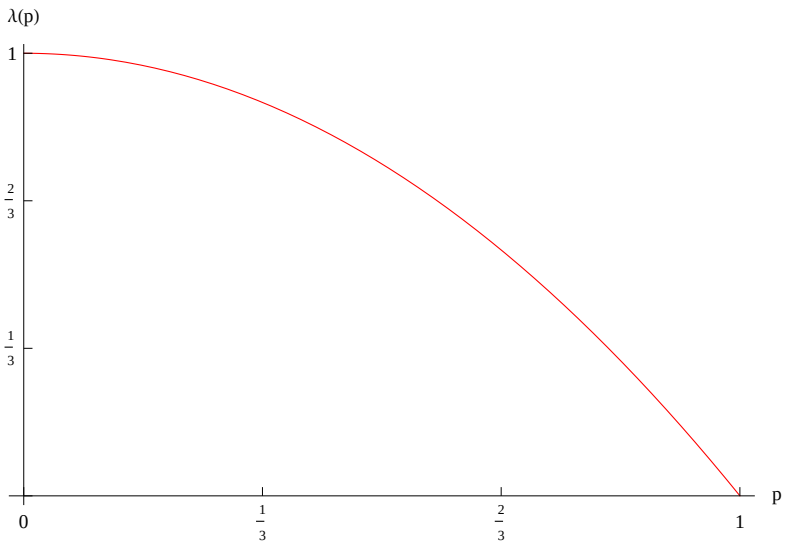


Figure 3.1: $p$ and $\lambda$ trade off against each other. In this example, $\lambda(p) = 1 - p^2$, satisfying assumptions 3.1, 3.2, 3.3, and 3.4.

to investigate incentives related to deliberate fraud, such as when data is misreported or fabricated, or when publication time is reduced through (self-) plagiarism. For formal work on credit-based incentives for fraud, see Bruner (2013) and Bright (forthcoming).

I make a number of assumptions on the way speed and reliability trade off against each other, as reflected in the speed function $\lambda$. The first assumption simply reflects the idea that they in fact trade off, as discussed above.

**Assumption 3.1** (The speed function is decreasing)**.** *The speed function* $\lambda : [0,1] \to \mathbb{R}$ *is such that for all* $p, p' \in [0,1]$*, if* $p < p'$*, then* $\lambda(p') < \lambda(p)$*.*

As I indicated above, assuming that $\lambda$ is decreasing simply reflects the fact that it takes more time to do research that is less likely to be erroneous. This is most obvious in the experimental sciences: collecting more data takes time.

**Assumption 3.2** (The speed function is concave)**.** *For every* $p, p', t \in [0,1]$*,*

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

This assumption may be described as a kind of decreasing marginal returns: as the reliability $p$ is lowered, the speed $\lambda$ is increased by assumption 3.1, but it increases ever slower as $p$ approaches zero: writing the papers itself takes time, which becomes relatively more significant if the scientist spends relatively little time on the research content. Conversely, if the scientist aims to be more reliable (increasing $p$), the speed $\lambda$ drops off ever faster. In other words, more and more extra research time is required to increase $p$ as $p$ approaches one.

**Assumption 3.3** (No perfect work)**.** $\lim_{p\to 1} \lambda(p) = 0$.

This assumption asserts that the scientist cannot deliver perfect work (in the sense of zero probability of errors), no matter how slowly she works. This reflects the fact that there is no certainty in science: for any fact or discovery, it is always possible that it will later be overturned, as Lakatos and Quine have argued.

**Assumption 3.4** (The speed function is differentiable)**.** *The function $\lambda$ is differentiable on the interior of its domain, i.e., for all $p \in (0, 1)$.*

This assumption says something about the "smoothness" of the tradeoff between speed and reliability. I have no opinion on whether this assumption is justified. In fact, it is hard to imagine what evidence for or against this assumption would even look like.[5] For this reason, when I state the results that can be proven based on these assumptions below, I consider both the case with and without assumption 3.4.

[5]When I have presented this paper, I have received extremely mixed responses to this. Some think that there are at least some cases where it would be important to model the function $\lambda$ as non-differentiable. Others insist that even if there are such cases, it should make no practical difference as any non-differentiable function can be approximated arbitrarily closely by a differentiable one. I prefer to maintain the agnostic stance I take in the main text.

How does all this affect the scientist's credit? For reasons I outlined above, I assume the scientist gets credit only for published work. Whether or not the scientist's work is published is determined through *peer review*. The purpose of peer review is to determine whether the results of the scientist's work are likely to stand up to the scrutiny of the scientific community.

In the simplest possible case it does this "pre-screening" perfectly: all and only those papers that are in fact accurate are accepted. The scientist does not know whether her paper is accurate; she only knows the reliability $p$, i.e., her credence that it is accurate. So from the scientist's perspective, if she produces a paper with reliability $p$, there is a probability $p$ that the journal publishes it.

Suppose that the amount of credit for a published accurate result is $c_a$. Then the scientist's expected credit per unit time is a function $C$ of the chosen reliability $p$ and the speed $\lambda$ (which is itself a function of $p$) given by $C(p) = c_a p\lambda(p)$.

In reality the peer review system cannot perfectly predict the future. Some accurate results get rejected, while some erroneous results get accepted. An example of the latter is Fleischmann and Pons' paper in the *Journal of Electroanalytical Chemistry*: it passed peer review but was thoroughly discredited within a year of publication (thus satisfying my definition of erroneous).

The acceptance of an erroneous result is called a false positive and the rejection of an accurate result a false negative. Following common usage in statistics I write $\alpha$ for the probability of a false positive and $\beta$ for the "power" (the probability that a false negative is avoided, i.e., that an accurate result is accepted). The case of "perfect peer review" described above would be one where $\beta = 1$ and $\alpha = 0$.

Here I assume instead that peer review is imperfect in the sense of a positive probability of false positives ($\alpha > 0$), or at least that the scientist believes this to be the case. Note that I remain agnostic on the possibility

of false negatives ($\beta$ may or may not be equal to one) although it seems reasonable to assume that those exist as well. I do assume that there is some discernment in the peer review system, i.e., accurate results are more likely to be accepted than erroneous results ($\beta > \alpha$).

**Assumption 3.5** (Imperfect peer review)**.** *The peer review acceptance probabilities $\alpha \in [0, 1]$ and $\beta \in [0, 1]$ are such that $\beta > \alpha > 0$.*

For the amount of credit for a published erroneous result I write $c_e$. While this reflects results that are "discredited", that does not necessarily equate to zero credit. Discredited research frequently still gets cited as if it was accurate (Tatsioni et al. 2007), even after a formal retraction (Budd et al. 1998). In other cases the fact that the proposed hypothesis has fallen out of favor does not prevent it from being a credit-worthy contribution to science, e.g., Priestley's work on phlogiston. This suggests that erroneous publications are worth some credit, although no more than accurate publications, i.e., $0 < c_e \leq c_a$.[6]

Putting all of this together yields the following. The scientist works on the research project at speed $\lambda(p)$. The result is accurate with (subjective) probability $p$. In this case it gets published with probability $\beta$ and this publication is worth $c_a$ units of credit. With probability $1 - p$ the result is erroneous, which leads to a publication worth $c_e$ units of credit with probability $\alpha$. Thus the scientist's expected credit per unit time, as a function of $p$, is given by

$$C(p) = c_a \beta p \lambda(p) + c_e \alpha (1 - p) \lambda(p).$$

To compare the individually optimal (i.e., credit-maximizing) tradeoff between speed and reliability to the socially optimal tradeoff, it is important

[6] On the other hand, some discredited research can actively harm a scientist's career (more so than publishing nothing at all would have done), suggesting that $c_e < 0$. These are usually cases of fraud rather than honest mistakes and so they are not my primary concern here. However, the point in the main text is not to argue that $c_e$ is necessarily positive, but that erroneous publications can influence a scientist's credit stock.

to be explicit about what is meant by the *social value* of scientific research. Here I have in mind the contribution that it makes to science as a social enterprise, which in turn benefits society. This is reflected in the first place by the extent to which the work is utilized by other scientists, and in the second place by the extent to which it or work based on it finds its way into society, say in the form of new technology.[7]

What is the social value $V$ of the scientist's research, as a function of her choice of reliability $p$? I assume that research can have social value only when it is published. The probabilities of publication $\alpha$ and $\beta$, the reliability $p$, and the speed $\lambda(p)$ are all as above.[8] The only difference is in the value of the research:

$$V(p) = v_a \beta p \lambda(p) + v_e \alpha (1-p) \lambda(p),$$

where $v_a$ is the social value of an accurate result, and $v_e$ the social value of an erroneous result.

Credit is awarded for (accurate) scientific work proportional to its social value, as has been argued in the literature. Merton enumerates the various kinds of rewards that exist in science—from the Nobel Prize to a journal publication—and concludes that "rewards are to be meted out in accord with the measure of accomplishment" (Merton 1957, p. 659). Strevens compares rewards in science to those given out in other areas and concludes that in

[7]Work in the natural sciences is perhaps more likely to produce new technology than work in the social sciences or the humanities. But all kinds of research can find its way into society. For example, research in economics may affect policy decisions, and philosophical research may affect the public debate.

[8]Hence the social value $V$ of the scientist's research is more precisely the scientist's own subjective estimate of the expected social value of the research (because $\alpha$, $\beta$, and $p$ are subjective probabilities). This may seem problematic when I use the function $V$ below to argue that credit incentivizes scientists to make choices that are not socially optimal. But as long as there are no structural biases (i.e., scientists' subjective probabilities roughly track the proportion of their work that turns out to be accurate) this presents no problems for the interpretation of my results, which are only concerned with averages anyway. Since in a competitive environment like science such biases would carry serious penalties, it seems reasonable to assume them away.

general "society accords prestige and other rewards... in proportion to the social good resulting from [the achievement]" (Strevens 2003, p. 78).

If a more exact measure of the amount of credit awarded to a specific publication (as opposed to a scientist) is wanted, a good candidate is the number of times it is cited. But at the same time the number of citations provides a measure of the extent to which the publication has been utilized by other scientists, which I argued reflects its social value. Based on these two lines of reasoning, I assume that $v_a = c_a$.

How about the social value of an erroneous result $v_e$? While errors can sometimes be instructive, I take it that the case in which they are distracting or actively misleading is more common. Take for instance a study which erroneously (in hindsight) finds a particular medicine helps cure some disease. Perhaps the error was in the design of the study, or perhaps it was simply bad luck, i.e., the data were acquired properly but they just happened to suggest a misleading conclusion. Either way, once the conclusion that the medicine is effective is published, it takes more time and effort to set the record straight than it would have to establish that the medicine is ineffective in the absence of the erroneous publication.[9] Moreover, before the error is corrected (and perhaps after as well, see Budd et al. 1998 and Tatsioni et al. 2007) the scientific community and society will proceed as if the medicine is effective, with potentially negative consequences for future research and public health.

So it seems to me that erroneous results are, on average, at best socially neutral, if not socially harmful: $v_e \leq 0$. However, I need not insist on this conclusion. For my purposes here it suffices that (a) the social value of erroneous results is less than the credit given for them ($v_e < c_e$), and (b) the social value of erroneous results is substantially less than that of accurate results ($v_e \leq v_a/2$). Both (a) and (b) follow immediately if my argument that $v_e \leq 0$ is accepted, but if one thinks that erroneous results have some

[9]Recall that I defined an erroneous result as one that does not hold up as true in the scientific community in the mid-term. Thus it is impossible by definition for an error to go uncorrected.

positive social value, my argument below still goes through as long as that value is relatively small.

Once again reasoning in terms of citations yields a similar conclusion. As mentioned above an erroneous result may still receive plenty of citations (Budd et al. 1998, Tatsioni et al. 2007). But here it does not seem so plausible that this a direct measure of its social value. Some of these citations may be actively criticizing the result. Others may be utilizing it under the assumption that it is accurate, possibly causing them to make errors in turn. This suggests that the social value of erroneous results is substantially less than that of accurate results, in support of (b).[10]

At the same time, citations to erroneous results are still worth credit, regardless of whether they are supportive or critical: they recognize the publication and its author as worth engaging with. The enormous effort undertaken by physicists to attempt to replicate Fleischmann and Pons' results (and when that failed to show that the mistake was with them, see Close 1991, chapter 12) is testament to Fleischmann and Pons' authority as competent electrochemists (Kitcher 1993, section 8.2). In contrast, work by subsequent cold fusion researchers has been largely ignored (Huizenga 1993, p. 208), as has work by creation scientists that challenges mainstream paleontology (Kitcher 1993, section 8.2). So the "credit value" of a citation to an erroneous publication is higher than its social value, in support of (a).

Assumption 3.6 summarizes what I have argued are reasonable constraints on the parameter values that reflect the credit and social value of scientific work in typical cases (including, in particular, the case that Fleischmann and Pons found themselves in).

**Assumption 3.6** (Credit and social value)**.**

*3.6.a. Accurate results have positive value:* $c_a > 0$ *and* $v_a > 0$.

[10]This argument assumes that erroneous results do not get cited more than accurate results. It seems unlikely that an erroneous result would, on average, get cited more than an accurate result. If it gets cited less, this provides further support for my conclusion. Tatsioni et al. (2007) provide some empirical support for this assumption.

*3.6.b. Accurate results are awarded credit proportional to their social value:* $c_a = v_a$.

*3.6.c. Erroneous results are awarded no more credit than accurate results:* $c_e \leq c_a$.

*3.6.d. The social value of erroneous results is less than the credit given for them:* $v_e < c_e$.

*3.6.e. The social value of erroneous results is at most half that of accurate results:* $v_e \leq v_a/2$.

I now state three results that can be proven based on the assumptions I have made. The first result states that the functions $C$ and $V$ have unique maxima, i.e., there is a particular reliability that a rational credit-maximizing scientist would choose, and there is a particular reliability that maximizes the social value of the scientist's contribution (which may be different or the same as the value that maximizes credit). This result does not use the controversial assumption 3.4 (which concerns the smoothness of the speed function) or the potentially controversial parts of assumption 3.6.

**Theorem 3.7.** *If assumptions 3.1, 3.2, 3.3, 3.5, and 3.6.a are satisfied, then there exist unique values* $p^*_C < 1$ *and* $p^*_V < 1$ *that maximize the functions* $C$ *and* $V$ *respectively, i.e.,*

$$C(p^*_C) = \max_{p\in[0,1]} C(p) \qquad \textit{and} \qquad V(p^*_V) = \max_{p\in[0,1]} V(p).$$

(See appendix B for proofs of the results in this section.)

Note that even with these fairly minimal assumptions, it follows that $p^*_V < 1$. This means that even from the social perspective perfect reliability is not a goal worth striving for. Or in other words, even if the scientist was "high-minded" in the sense that she only cared about maximizing the social value of her scientific work, she should not strive to avoid error at all cost.

That $p_V^* < 1$ is a more or less direct consequence of assumption 3.3 ("no perfect work") and hence reflects the insight of Lakatos and Quine that there is no certainty in science. It means that even in a science functioning perfectly, a tradeoff between speed and reliability must be made, and hence errors should be expected. This reflects back on the discussion of peer review: it is designed on the basic premise that there will be errors, and science must attempt to catch them as early as possible. If errors are expected, it also seems wrong to hold it against individual scientists too much when they make mistakes, which gives some additional philosophical support for assumption 3.6.d ($v_e < c_e$).

The second result says that the imperfections in the peer review system and the way credit is awarded systematically favor lower levels of reliability. That is, a scientist who maximizes expected credit will choose a reliability no higher than the optimal level from the perspective of maximizing social value. This result does not use the controversial assumption 3.4 or the assumption that the social value of erroneous results is at most half that of accurate results (assumption 3.6.e).

**Theorem 3.8.** *Let assumptions 3.1, 3.2, 3.3, 3.5, and 3.6.a–3.6.d be satisfied, and define $p_C^*$ and $p_V^*$ as in theorem 3.7. Then $p_C^* \leq p_V^*$.*

I interpret this result as showing that, given imperfect peer review, there is a credit-incentive to produce research at a systematically lower reliability than is socially optimal. But it could easily be objected that I have not shown this: theorem 3.8 leaves open the possibility that $p_C^* = p_V^*$, the happy case in which individual and social incentives align exactly.

However, the happy case can only arise in one of two situations. First, if the value of erroneous results is so high that it is both individually and socially optimal to have no concern whatsoever for reliability ($p_C^* = p_V^* = 0$, see figure 3.2). Second, if the speed function is not differentiable at the point of optimality (see figure 3.3). The third result shows that if these two

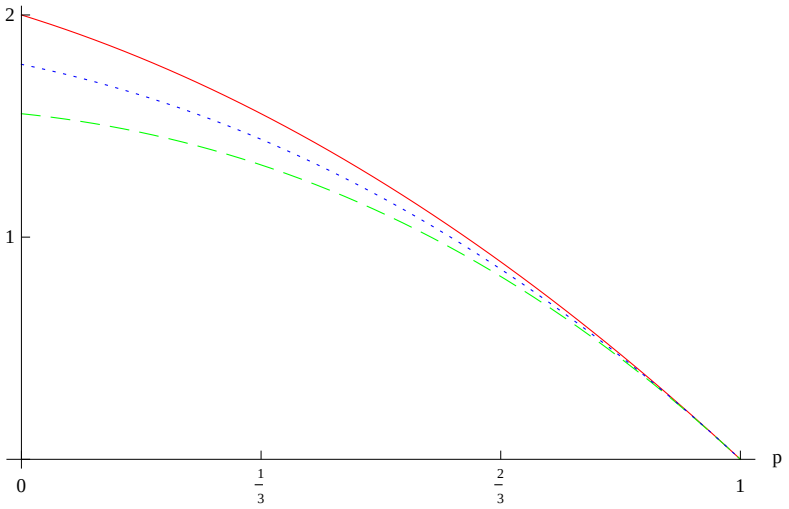


Figure 3.2: If $\lambda(p) = 2 - p - p^2$ (the solid red line) and $v_e$ is relatively high, it may be that $p_C^* = p_V^* = 0$. In this example, the function $C$ is shown as a dotted blue line (with $c_a\beta = 1$ and $c_e\alpha = 8/9$) and the function $V$ is shown as a dashed green line (with $v_a\beta = 1$ and $v_e\alpha = 7/9$).

situations are ruled out there is a definite misalignment between individual and social incentives.

**Theorem 3.9.** *Let assumptions 3.1, 3.2, 3.3, 3.4, 3.5, and 3.6 be satisfied, and define $p_C^*$ and $p_V^*$ as in theorem 3.7. Then $p_C^* < p_V^*$.*

So when all the assumptions are brought into play, credit-maximization gives the scientist an incentive to work faster, but less reliably, than is optimal from the perspective of maximizing social value.[11]

This result depends crucially on the imperfections in the peer review system, and in particular the possibility of false positives. If $\alpha = 0$ and

[11]Some of the assumptions can be weakened: I have chosen to present them in a way that allows me to focus on their plausibility, rather than stating them as generally as possible. For example, in theorems 3.7 and 3.8 assumption 3.5 (imperfect peer review) can be weakened: $\beta > 0$ and $\beta \geq \alpha$ suffices for theorem 3.8, and just $\beta > 0$ suffices for theorem 3.7. In theorems 3.8 and 3.9 assumption 3.6.b ($v_a = c_a$) can be weakened: if $c_e \geq 0$ then $v_a \geq c_a$ suffices and if $c_e \leq 0$ then $v_a \leq c_a$ suffices.

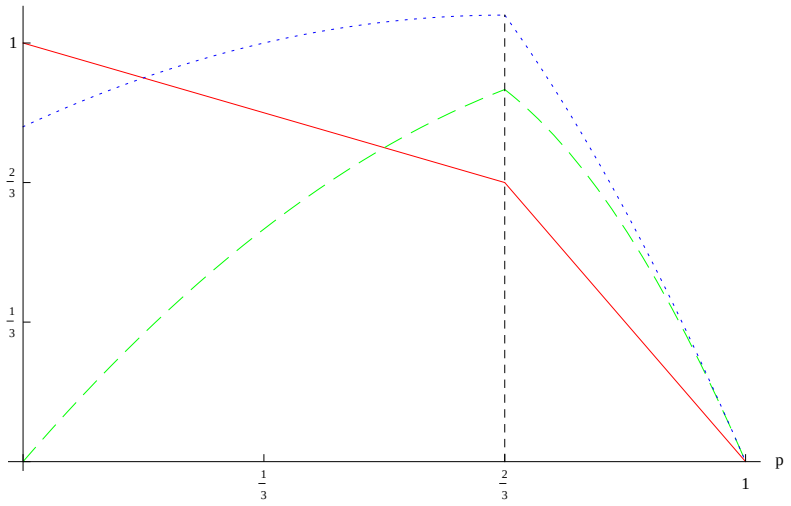


Figure 3.3: If $\lambda(p) = 1 - p/2$ for $p \leq 2/3$ and $\lambda(p) = 2(1-p)$ for $p > 2/3$ then $\lambda$ (solid red) is not differentiable at $p = 2/3$. Then the functions $C$ (dotted blue, with $c_a\beta = 2$ and $c_e\alpha = 4/5$) and $V$ (dashed green, with $v_a\beta = 2$ and $v_e\alpha = 0$) may both be maximized there: $p^*_C = p^*_V = 2/3$

$\beta > 0$ then assumptions 3.1, 3.2, 3.3, and 3.6.a are sufficient to show that the functions $C$ and $V$ have unique maxima, and that these maxima are equal. Intuitively, given imperfect peer review it makes sense for scientists to quickly produce lots of papers and "see what sticks" rather than spending too much time perfecting any one paper.

Hence I interpret theorems 3.8 and 3.9 as showing that imperfections in the peer review system create a systematic bias that leads credit-maximizing scientists to favor speed over reliability relative to the social optimum. While individual scientists may have other goals than maximizing credit, this means that scientists have an incentive to rush into print: the way their work is rewarded with credit systematically favors a focus on speed over reliability.

## 3.4 A Tradeoff Between Speed, Reliability, and Impact

One feature of Fleischmann and Pons' work that presumably played a role in their decision to go public but did not appear in the model so far is the potential *impact* of their work. As the media emphasized in the days after the press conference, if cold fusion worked it held the promise of an energy revolution.

Fleischmann and Pons could perhaps be described as "impact-seekers", scientists who go in for risky research in relatively unexplored areas that promises to yield great rewards if successful (Close 1991, p. 71, describes Fleischmann in this way). In contrast, many scientists are "safety-seekers", content to make small contributions that are likely to be correct and accepted and/or can be made relatively quickly. The distinction is analogous to that between mavericks and followers (Weisberg and Muldoon 2009) or explorers and extractors (Thoma 2015), but goes back much further (e.g., Hull 1988, p. 474). I use different terminology to avoid the implication that this is a binary distinction rather than a graded one, or that there is necessarily a psychological explanation for it.

In this section I expand the previous model to include research with differential potential impact. The scientist now has to make a three-way tradeoff. She chooses both the reliability and the impact, but choosing either or both of these too highly comes at the expense of speed (compare the old business saying "You can have it good, fast, or cheap; pick two").

Above I showed that if there are imperfections in the peer review system, scientists tend to favor speed over reliability (relative to the social optimum). The first question I aim to investigate here is what the consequences of imperfections in the peer review system are in this more complicated model. The second question is to what extent the different "types" of scientists—impact-seekers and safety-seekers—show up in the model. More specifically,

can credit incentives explain the existence of both types?[12]

In the model of this section the scientist chooses both a desired reliability $p$ and a desired impact level $c$. Since $p$ is interpreted as a probability, its domain is naturally constrained to the interval $[0, 1]$. The impact $c$ is not similarly constrained. However, I assume that, at least for a given value of $p$, there is a maximum impact that can be achieved $\mu(p)$. For any admissible choice of $p$ and $c$, $\lambda(p, c)$ gives the scientist's speed. The following definitions formalize this setup.

**Definition 3.10.** The *maximum impact function* is a function $\mu : [0, 1] \to [0, \infty)$. The *set of admissible choices* is the set $D = \{(p, c) \mid p \in [0, 1], c \in [0, \mu(p)]\}$. The *speed function* has $D$ as its domain: it is a function $\lambda : D \to \mathbb{R}$.

I make a number of assumptions on the shape of $\lambda$. These assumptions are very similar to the ones I made before. Although they have to be adapted to the new context, their justification is as before.

First I assume that the speed function is decreasing in each of its arguments. That is, at a fixed level of reliability, increasing the impact decreases speed, and at a fixed level of impact, increasing reliability decreases speed.

**Assumption 3.11** (The speed function is decreasing)**.**

*3.11.a. For all $p, p' \in [0, 1]$, if $p < p'$ and $c \leq \min\{\mu(p), \mu(p')\}$, then $\lambda(p', c) < \lambda(p, c)$.*

*3.11.b. For all $p \in [0, 1)$, if $c < c' \leq \mu(p)$, then $\lambda(p, c') < \lambda(p, c)$.*[13]

**Assumption 3.12** (The speed function is concave)**.** *For any $(p, c), (p', c') \in D$ and $t \in [0, 1]$,*

[12] This question is raised by Thoma (2015, section 4.4). She points out that, from the purely epistemic perspective taken by Weisberg and Muldoon (2009), this cannot be explained: "In their model, it was unclear why anybody would choose to be a [safety-seeker], given their lack of productivity. In [Thoma's model], the question is why anybody would choose to be an [impact-seeker]" (Thoma 2015, p. 470).

[13] This assumption excludes the case where $p = 1$. This is because subsequent assumptions entail that $\lambda(1, c) = 0$ for all $c$, which would contradict this assumption if $\mu(1) > 0$.

*3.12.a.* $(tp + (1-t)p', tc + (1-t)c') \in D$;[14]

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

As before, this assumption says that there are decreasing marginal returns from decreasing reliability to gain speed. This more general version says that there are also decreasing marginal returns from decreasing the impact level, which is justified for the same reason.

**Assumption 3.13.** *The function $\lambda$ vanishes as $p$ or $c$ approaches the edge of its domain $D$.*

*3.13.a.* $\lim_{p\to 1} \lambda(p,0) = 0$.

*3.13.b. For all* $p \in [0,1]$, $\lim_{c\to\mu(p)} \lambda(p,c) = 0$.

This assumption has a role similar to assumption 3.3 (the "no perfect work" assumption). Assumption 3.13.a is in fact identical to that assumption (although for technical reasons I only need to make the assumption for the case $c = 0$) and has the same justification. Assumption 3.13.b formalizes the idea that $\mu(p)$ represents the maximum impact that can be achieved at a given reliability $p$, by requiring that the scientist's speed becomes negligible as this value is approached.

**Assumption 3.14** (The speed function is differentiable (in $p$))**.** *The partial derivative of the function $\lambda$ with respect to its first argument exists on the interior of its domain, i.e., $\frac{\partial}{\partial p}\lambda(p,c)$ exists whenever $0 < p < 1$ and $0 < c < \mu(p)$.*

This assumption requires that the speed function is "smooth" (at least in one direction). As before, I consider results both with and without this assumption.

[14]It does not follow from the definition of the domain $D$ of $\lambda$ or the assumptions made so far that $(tp + (1-t)p', tc + (1-t)c') \in D$, but this is required for the idea of a concave function to make sense, hence this assumption. It is equivalent to the assumption that $\mu$ is a concave function.

What does the credit function look like in this more general setting? The main difference is that the credit for an accurate result is no longer an exogenously fixed parameter $c_a$, but a variable $c$ whose value is chosen by the scientist. As for the credit for an erroneous result, there is a modeling choice to be made. Either it is a fixed absolute value, independent of the impact the result would have had if it was accurate, or it is proportional to the impact. Here I choose the latter option (although I suspect that similar results could be proven if the former option was used).

So credit for erroneous results ($c_e$ in the previous section) is now given by $r_c c$: a proportionality constant $r_c$ times the scientist's chosen impact level $c$. If $r_c > 0$, this means that erroneous results that would have had a high impact get more credit ("at least you tried something ambitious"). If $r_c < 0$, this means that erroneous results of potentially high impact are penalized more harshly ("the bigger they are, the harder they fall"). This seems right at least for the case of Fleischmann and Pons: the amount of attention given to proving them wrong, and the effect on their personal reputations, seems to have been bigger exactly because of the potential impact their work could have had.

So the scientist's expected credit, as a function of $p$ and $c$, is

$$C(p, c) = \beta pc\lambda(p, c) + \alpha(1 - p)r_c c\lambda(p, c).$$

Now consider the social value of the scientist's work. I assume that the impact level $c$ chosen by the scientist reflects not only the potential reward (credit) but also the potential social value of the work. So the variable $c$ replaces not only the parameter $c_a$ but also the parameter $v_a$. This is equivalent to assumption 3.6.b ($c_a = v_a$) but for notational convenience here I build this assumption into the definition of the function $V$ rather than stating it separately.

As I did for the case of credit, I assume that the social value of an erroneous result is determined in proportion to the value of an accurate result,

i.e., $v_e$ is replaced by $r_v c$, where $r_v$ is the proportionality constant for the social value of erroneous results. So the social value of the scientist's research, as a function of $p$ and $c$, is

$$V(p, c) = \beta pc\lambda(p, c) + \alpha(1 - p)r_v c\lambda(p, c).$$

The assumption on the peer review parameters $\alpha$ and $\beta$ is exactly like before. I restate it here as a reminder.

**Assumption 3.15** (Imperfect peer review)**.** *The peer review acceptance probabilities $\alpha \in [0, 1]$ and $\beta \in [0, 1]$ are such that $\beta > \alpha > 0$.*

The assumptions on the parameters $r_c$ and $r_v$ are similar to the assumptions on $c_e$ and $v_e$ I made in assumption 3.6, with the following changes. First, assumptions 3.6.a and 3.6.b are no longer needed because they are built into the definition of the functions $C$ and $V$. Second, assuming that $r_c \leq 1$ is interpretationally equivalent to the assumption that $c_e \leq c_a$ which I made above (because $r_c \leq 1$ if and only if $r_c c \leq c$). Third, for technical reasons, a slightly stronger assumption on the value of erroneous results is needed: $r_v \leq 1/3$ instead of $v_e \leq v_a/2$.

**Assumption 3.16** (Credit and social value)**.**

*3.16.a. Erroneous results are awarded no more credit than accurate results: $r_c \leq 1$.*

*3.16.b. The social value of erroneous results is less than the credit given for them: $r_v < r_c$.*

*3.16.c. The social value of erroneous results is at most a third that of accurate results: $r_v \leq 1/3$.*

I present three results that use some or all of the above assumptions. The first result says that there are unique choices of reliability and impact level that maximize expected credit and that maximize social value. Assumptions 3.14 and 3.16 are not needed.

**Theorem 3.17.** *If assumptions 3.11, 3.12, 3.13, and 3.15 are satisfied, then there exist unique points* $(p_C^*, c_C^*)$ *and* $(p_V^*, c_V^*)$ *that maximize the functions* $C$ *and* $V$ *respectively, i.e.,*

$$C(p_C^*, c_C^*) = \max_{(p,c)\in D} C(p,c) \qquad \textit{and} \qquad V(p_V^*, c_V^*) = \max_{(p,c)\in D} V(p,c).$$

*Moreover,* $p_C^* < 1$ *and* $0 < c_C^* < \mu(p_C^*)$*; and* $p_V^* < 1$ *and* $0 < c_V^* < \mu(p_V^*)$.

(See appendix C for proofs of the results in this section.)

If assumptions 3.16.a and 3.16.b are added the credit-maximizing reliability $p_C^*$ is at most the social value maximizing reliability $p_V^*$.

**Theorem 3.18.** *Let assumptions 3.11, 3.12, 3.13, 3.15, 3.16.a, and 3.16.b be satisfied, and define* $(p_C^*, c_C^*)$ *and* $(p_V^*, c_V^*)$ *as in theorem 3.17. Then* $p_C^* \leq p_V^*$.

And, finally, if assumptions 3.14 and 3.16.c are added the inequality is strict.

**Theorem 3.19.** *Let assumptions 3.11, 3.12, 3.13, 3.14, 3.15, and 3.16 be satisfied, and define* $(p_C^*, c_C^*)$ *and* $(p_V^*, c_V^*)$ *as in theorem 3.17. Then* $p_C^* < p_V^*$.

How do these results shed light on the two questions I raised above?

First, imperfections in the peer review system give the scientist an incentive to favor speed and/or impact over reliability, relative to what she would do if she were trying to maximize the social value of her work. This is true under essentially the same conditions as above. So the results expressed in theorems 3.8 and 3.9 are seen to be robust against the introduction of the dimension of impact.

Second, theorem 3.17 rules out the possibility that a scientist could switch from being a safety-seeker to an impact-seeker (increasing impact at the expense of reliability) or vice versa, while remaining at a global maximum of either $C$ or $V$. For a credit-maximizing scientist, there is just one rational

choice, 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.

This does not rule out the existence of different "types" of scientists. But it suggests that these types are the result of differences in the shape of the speed function of different scientists. If the speed function describes the tradeoff between reliability, impact, and speed for a given scientist, the location of the optimum given that particular speed function determines the type of scientist she will be (or at least has a credit-incentive to be). If the speed function is more or less fixed over the course of a career[15] and outside the scientist's control, theorem 3.17 can be interpreted as showing that different types of scientists are the result of differences in aptitude rather than choice.

Whether a scientist is likely to be an impact-seeker or a safety-seeker is thus determined by the shape of her speed function. The following example illustrates this.

**Example 3.20.** Consider two scientists. For scientist 1, the tradeoff between reliability, impact, and speed is given by the speed function $\lambda_1$, where

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

for all $0 \leq p \leq 1$ and $0 \leq c \leq \frac{1}{2}(\sqrt{25 + 12p - 12p^4} - 3 - 2p)$ (see figure 3.4). Note that this function satisfies assumptions 3.11, 3.12, 3.13, and 3.14. Suppose further that $r_c = 0$.[16] Then the credit-maximizing choice for scientist 1 is $p \approx 0.52$ and $c \approx 0.38$.

In contrast, scientist 2's speed function is given by

[15] See Huber (2001, and citations therein) for evidence that the productivity of scientists is, on average, constant over the course of a career.

[16] As a result I need to make no specific assumption on the values of $\alpha$ and $\beta$: the maximum of $C$ will not depend on this as long as $\beta > 0$.

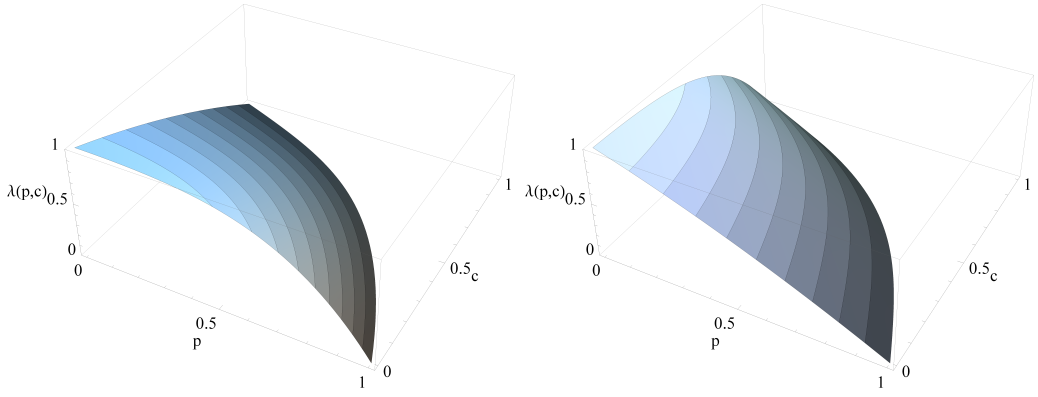


Figure 3.4: Graphs of $\lambda_1$ (on the left) and $\lambda_2$ (on the right).

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

for all $0 \leq c \leq 1$ and $0 \leq p \leq \frac{1}{2}(\sqrt{25 + 12c - 12c^4} - 3 - 2c)$ (see figure 3.4). This function also satisfies assumptions 3.11, 3.12, 3.13, and 3.14. But the credit-maximizing choice for scientist 2 is $p \approx 0.38$ and $c \approx 0.52$.

Scientist 1's speed function "favors" reliability compared to scientist 2's, which "favors" impact. This is because $\lambda_1$ is closer to linear in $c$—having only a small quadratic component—while it is a fourth-degree polynomial in $p$ ($\lambda_2$ 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.

In general, given that the speed function is concave and decreasing, the less linear it is in one of its variables the higher the optimal value for that variable will be. So more linear behavior in $c$ and less linear behavior in $p$ produces safety-seekers, and the reverse impact-seekers.

Moreover, if scientists are credit-maximizers, and assumptions 3.11, 3.12, 3.13, and 3.15 are justified, then theorem 3.17 guarantees that differences in the shape of the speed function are the *only* way different types of scientists

can arise.

This is potentially a problem. Weisberg and Muldoon (2009) and Thoma (2015) have investigated whether there is an epistemically optimal distribution of safety-seekers and impact-seekers in a scientific community. Thoma (2015, section 4.4) points out that credit may play a crucial role in motivating scientists to distribute themselves over the types. But if differences in aptitude are required for credit to play this role, there is no reason to expect the resulting distribution of safety-seekers and impact-seekers to be anywhere close to optimal.

## 3.5 Conclusion

The following four conclusions can be drawn from the work presented in this paper.

First, imperfections in the peer review system create a misalignment between when it is optimal to "go public" from a credit-maximizing perspective and the socially optimal time to do so. This misalignment systematically sacrifices reliability. So scientists have a credit-incentive to produce work that is less reliable than is socially optimal. This is true when only the tradeoff between speed and reliability is considered as well as when a three-way tradeoff between speed, reliability, and impact is considered (in each case, under some plausible assumptions on the way they trade off).

This misalignment hurts science and society: by definition, any deviation from the social optimum hurts the progress of science and the social benefits of that progress. More specifically, lower reliability implies that more errors enter the scientific literature. The combination of imperfections in the peer review system and credit-maximization is thus a source of reproducibility problems.

What can be done about this? One solution is to eliminate imperfections in the peer review system. Without those imperfections credit-incentives are

perfectly aligned with the social optimum in my model. But this is a lot to ask: it requires reviewers at scientific journals to never make mistakes in predicting the reception of the paper by other scientists in the short- to mid-term.

However, I noted that the misalignment of incentives in the model is exclusively caused by false positives (accepting erroneous results for publication). So reducing those can bring the credit-maximizing optimum closer to the social optimum.[17] This would seem to recommend conservative editorial practices: rejecting papers even based on fairly minimal doubts about their accuracy. But if reducing false positives leads to more false negatives (rejecting accurate results) the effect will be that the maximum social value is itself lowered, even if the credit-maximizing optimum is brought closer to it. Investigating that particular tradeoff is beyond the scope of this paper.

A different way to eliminate imperfections in the peer review system would be to get rid of peer review (and perhaps even scientific journals) altogether. But even such a drastic rethinking of the way scientific research is disseminated would not avoid this problem. The problem arises because scientific work needs to be evaluated in some way or other in the short run, whereas its accuracy is not known with certainty until at least the mid-term. A system that could predict accuracy perfectly seems in principle impossible, as it would need to predict the outcomes of future research that establishes the accurate or erroneous nature of the present work. Hence, while I have focused discussion on imperfections in the peer review system, the existence of peer review in its current form is not essential to the problem.

Another solution would focus on the amount of credit given for erroneous results. I referred repeatedly to Budd et al. (1998) and Tatsioni et al. (2007), who showed that scientists continue to give credit (in the form of citations) to research that has been shown to be erroneous. If the credit given to erroneous

[17] More specifically, it can be shown (under the assumptions of theorems 3.9 and 3.19) that reducing the value of $\alpha$ reduces the difference between $p_C^*$ and $p_V^*$. In the limiting case where $\alpha = 0$ they are equal.

results matched the social value of those results more closely, the gap between the credit-maximizing optimum and the social optimum would be reduced. It would be helpful, for example, if there was a broader awareness of which research has been shown to be erroneous. But again this may be hard to achieve in practice.

A third solution would be to try to somehow compensate for the misalignment. For example, Nelson et al. (2012) have suggested limiting the number of papers scientists may publish per unit time. This would create an incentive to favor reliability over speed that could in principle balance out the misalignment I have shown. But this suggestion comes with its own problems. The limit on the number of papers would have to be just right to balance out the incentive to favor speed over reliability without overshooting the optimum in the other direction, needlessly harming the timely publication of accurate results. This problem is exacerbated by the fact that different scientists may have different speed functions, which may require different publication limits to create the best incentive structure.

As all of these suggestions have some problems associated with them, it is not clear which one(s) should be recommended. I leave a more detailed comparison of these and other possible solutions to future work.

The second conclusion is that perfect reliability is neither to be expected nor to be desired. The reason for this is of course that if scientists were too demanding in perfecting their research before publishing it, nothing would ever get published. The point is hardly new (it goes back at least to Lakatos and Quine), but since philosophers of science and epistemologists have said a lot about error avoidance but relatively little about how to achieve this in a reasonable time frame (cf. Friedman 1979, Heesen 2015), it is worth emphasizing.

Third, I considered the difference between scientists who pursue high-impact research that is risky with regard to reliability and/or speed ("impact-seekers") and scientists who pursue more mundane research relatively likely

to be accurate and/or fast ("safety-seekers"). My model suggests that the existence of these types of scientists reflects a difference in aptitude rather than a preference for certain kinds of research: impact-seekers are scientists with an aptitude for high-impact research at a relatively small cost in speed, while safety-seekers can pursue highly reliable research at a relatively small cost in speed.

Considering the tradeoff between speed, reliability, and impact explicitly shows that high-impact research (or "transformative" research in modern terms) is likely to be less reliable. Example 3.20 illustrates this. Thus it seems unreasonable to hold impact-seekers to the same standards of evidence as safety-seekers. In this way my model justifies to some extent the practice at institutions like the NSF and the NIH to consider a grant proposal's "potential to be transformative" separately from its likelihood to succeed. By considering the criteria separately, these institutions aim to prevent biasing their evaluation process for or against impact-seekers or safety-seekers.

Finally, the work in this paper suggests a reevaluation of Fleischmann and Pons' decision to go public with their work on cold fusion. That decision has been much maligned for being premature. The rejection of cold fusion by the scientific establishment and the subsequent decline of cold fusion research would seem to vindicate the judgment of prematurity. But Fleischmann and Pons could not know this at the time. The question is whether their decision was irrational, given the information available to them.

Two of the above conclusions suggest that it may not have been. First, imperfections in the peer review system may make it rational for a credit-maximizing scientist to submit relatively unreliable work, i.e., work with a relatively high chance of later being proven wrong. Second, scientists who are pursuing high-impact research should be given more leeway to produce relatively unreliable results.

Fleischmann and Pons were well aware of the uncertainties surrounding cold fusion at the time they went public. They also knew that if they did not

go public, the risk of being scooped was extremely high. The above considerations suggest (without proving of course) that under these circumstances it may well have been rational to go public despite the uncertainties.

Fleischmann and Pons went out on a limb, as every scientist does when she publishes her work. On this occasion, they got burned. But I submit that this was not primarily the result of poor judgment, although it may be easy to come to the opposite conclusion with the benefit of hindsight. Rather, they did exactly what other scientists have done on countless occasions: they weighed the risk of going public against the potential reward. That they are now maligned rather than celebrated is largely the result of bad luck.

# Chapter 4

# When Journal Editors Play Favorites

## 4.1 Introduction

Journal editors occupy an important position in the scientific landscape. By making the final decision on which papers get published in their journal and which papers do not, they have a significant influence on what work is given attention and what work is ignored in their field (Crane 1967).

In this paper I investigate the following question: should the editor be informed about the identity of the author when she is deciding whether to publish a particular paper? Under a single- or double-blind reviewing procedure, the editor has access to information about the author, whereas under a triple-blind reviewing procedure she does not. So in other words the question is: should journals practice triple-blind reviewing?

Two kinds of arguments have been given in favor of triple-blind reviewing. One focuses on the treatment of the author by the editor. On this kind of argument, revealing identity information to the editor will lead the editor to (partially) base her judgment on irrelevant information (such as the gender of the author, or whether or not the editor is friends with the author). This

harms the author, and is thus bad.

The second kind of argument focuses on the effect on the journal and its readers. Again, the idea is that the editor will base her judgment on identity information if given the chance to do so. But now the further claim is that as a result the journal will accept worse papers. After all, if a decision to accept or reject a paper is influenced by the editor's biases, this suggests that a departure has been made from a putative "objectively correct" decision. This harms the readers of the journal, and is thus bad.

Here I provide a philosophical discussion of the reviewing procedure to assess these arguments. I distinguish between two different ways the editor's judgment may be affected if the author's identity is revealed to her. First, the editor may treat authors she knows differently from authors she does not know. Second, the editor may treat authors differently based on their membership of some group (e.g., gender bias). My discussion focuses on the following three claims.

My first claim is that the first kind of differential treatment the editor may display (based on whether she knows a particular author) actually benefits rather than harms the readers of the journal. This benefit is the result of a reduction in editorial uncertainty about the quality of submitted papers when she knows their authors. I construct a model to show in a formally precise way how such a benefit might arise—surprisingly, no assumption that the scientists the editor knows are somehow "better scientists" is required—and I cite empirical evidence that such a benefit indeed does arise. However, this benefit only applies in certain fields. I argue that in other fields (in particular, mathematics and the humanities) no significant reduction of uncertainty—and hence no benefit to the readers—occurs (section 4.2).

My second claim is that either kind of differential treatment the editor may display (based on whether she knows authors or based on bias against certain groups) harms authors. I argue that any instance of such differential treatment constitutes an epistemic injustice in the sense of Fricker (2007)

against the disadvantaged author. If the editor is to be (epistemically) just, she should prevent such differential treatment, which can be done through triple-blind reviewing. So I endorse an argument of the first of the two kinds I identified above: triple-blind reviewing is preferable because not doing so harms authors (section 4.3).

My third claim is that whether differential treatment also harms the journal and its readers depends on a number of factors. Differential treatment by the editor based on whether she knows a particular author may benefit readers, whereas differential treatment based on bias against certain groups may harm them. Whether there is an overall benefit or harm depends on the strength of the editor's bias, the relative sizes of the different groups, and other factors, as I illustrate using the model. As a result I do not in general endorse the second kind of argument, that triple-blind reviewing is preferable because readers of the journal are harmed otherwise. However, I do endorse this argument for fields like mathematics and the humanities, where I claim that the benefits of differential treatment (based on uncertainty reduction) do not apply (section 4.4).

Note that, in considering the ethical and epistemic effects of triple-blind reviewing, a distinction is made between the effects on the author and the effects on the readers of the journal. This reflects a growing understanding that in order to study the social epistemology of science, what is good for an individual inquirer must be distinguished from what is good for the wider scientific community (Kitcher 1993, Strevens 2003, Mayo-Wilson et al. 2011).

Zollman (2009) has studied the effects of different editorial policies on the number of papers published and the selection criteria for publication, but he does not focus specifically on the editor's decisions and the uncertainty she faces. Economists have studied models in which editor decisions play an important role (Ellison 2002a, Faria 2005, Besancenot et al. 2012), but they have not distinguished between papers written by scientists the editor knows and papers by scientists unknown to her, and neither have they been con-

cerned with biases the editor may be subject to. And some other economists have done empirical work investigating the differences between papers with and without an author-editor connection (Laband and Piette 1994a, Medoff 2003, Smith and Dombrowski 1998, more on this later), but they do not provide a model that can explain these differences. This paper thus fills a gap in the literature.

## 4.2 A Model of Editor Uncertainty

As I said in the introduction, journal editors have a certain measure of power in a scientific community because they decide which papers get published.[1] An editor could use this discretionary power to the benefit of her friends or colleagues, or to promote certain subfields or methodologies over others. This phenomenon has been called *editorial favoritism*. If anecdotal evidence is to be believed, this phenomenon is widespread. Some systematic evidence of favoritism exists as well. Bailey et al. (2008a,b) find that academics believe editorial favoritism to be fairly prevalent, with a nonnegligible percentage claiming to have perceived it firsthand. Laband (1985) and Piette and Ross (1992) find that, controlling for citation impact and various other factors, papers whose author has a connection to the journal editor are allocated more journal pages than papers by authors without such a connection.[2]

In this paper, I refer to the phenomenon that editors are more likely to accept papers from authors they know than papers from authors they do not know as *connection bias*.

[1] Different journals may have different policies, such as one in which associate editors make the final decision for papers in their (sub)field. Here, I simply define "the editor" to be whomever makes the final decision whether to publish a particular paper.

[2] Here, page allocation is used as a proxy for journal editors' willingness to push the paper. The more obvious variable to use here would be whether or not the paper is accepted for publication. Unfortunately, there are no empirical studies which measure the influence of a relationship between the author and the editor on acceptance decisions directly. Presumably this is because information about rejected papers is usually not available in these kinds of studies.

Academics tend to disapprove of this behavior (Sherrell et al. 1989, Bailey et al. 2008a,b). In both of the studies by Bailey et al., in which subjects were asked to rate the seriousness of various potentially problematic behaviors by editors and reviewers, this disapproval was shown (using a factor analysis) to be part of a general and strong disapproval of "selfish or cliquish acts" in the peer review process. Thus it appears that the reason for the disapproval of editors publishing papers by their friends and colleagues is that it shows the editor acting on private interests, rather than displaying the disinterestedness that is the norm in science (Merton 1942).

On the other hand, if connection bias was a serious worry for authors, one would expect this to be a major consideration for them in choosing where to submit their papers (i.e., submit to journals where they know the editor), but Ziobrowski and Gibler (2000) find that this is not the case.[3] Moreover, despite working scientists' disapproval, there is some evidence that connection bias improves the overall quality of accepted papers (Laband and Piette 1994a, Medoff 2003, Smith and Dombrowski 1998). Does that mean scientists are misguided in their disapproval?

As indicated in the introduction, I distinguish between the effects of editors' biases on the authors of scientific papers on the one hand, and the effects on the readers of scientific journals on the other hand. In this section, I use a formal model to show that these two can come apart: connection bias may negatively affect scientists as authors while positively affecting scientists as readers. Note that in this section I focus only on connection bias.

[3]In particular, authors who know an editor and thus could expect to profit from connection bias would find knowing the editor and the composition of the editorial board more generally to be important factors in deciding where to submit, contrary to Ziobrowski and Gibler's evidence (these factors are ranked twelfth and sixteenth in importance in a list of sixteen factors that might influence the decision where to submit). Similarly, authors who do not know an editor would find a lack of (perceived) connection bias and the composition of the editorial board to be important factors, but these rank only seventh and twelfth in importance in Ziobrowski and Gibler's study. In a similar survey by Mackie (1998, chapter 4), twenty percent of authors indicated that knowing the editor and/or her preferences is an important consideration in deciding where to submit a paper.

Subsequent sections consider other biases.

Consider a simplified scientific community consisting of a set of scientists. Each scientist produces a paper and submits it to the community's only journal which has one editor.

Some papers are more suitable for publication than others. I assume that this suitability for publication can be measured on a single numerical scale. For convenience I call this the *quality* of the paper. However, I remain neutral on how this notion should be interpreted, e.g., as an objective measure of the epistemic value of the paper (which is perhaps an aggregate of multiple relevant criteria), or as the number of times the paper would be cited in future papers if it was published, or as the average subjective value each member of the scientific community would assign to it if they read it.[4]

Crucially, the editor does not know the quality of the paper at the time it is submitted. The aim of this section is to show how uncertainty about quality can lead to connection bias. To make this point as starkly as possible, I assume that the editor cares only about quality, i.e., she makes an estimate of the quality of a paper and publishes those and only those papers whose quality estimate is high.

Let $q_i$ be the quality of the paper submitted by scientist $i$. Since there is uncertainty about the quality, $q_i$ is modeled as a random variable. Since some scientists are more likely to produce high quality papers than others, the mean $\mu_i$ of this random variable may be different for each scientist. I assume that quality follows a normal distribution with fixed variance: $q_i \mid \mu_i \sim N(\mu_i, \sigma^2_{qu})$.

The assumptions of normality and fixed variance are made primarily to keep the mathematics simple. Below I make similar assumptions on the distribution of average quality in the scientific community and the distribution of reviewers' estimates of the quality of a paper. There is no particular rea-

[4]For more on potential difficulties with interpreting the notion of quality, see Bright (2015).

son to expect quality to follow a normal distribution. One might also expect variances to vary, for example, particularly "good" scientists might not only have a high average quality, but might also be more reliable in the sense of having a lower variance around this mean. Or there might be different types of scientists: some who produce high-quality papers consistently, and others whose average is not particularly high but who occasionally produce a paper of spectacularly high quality. Whether and under what circumstances departures from the assumptions of normality and fixed variance lead to different results is a question I leave for future research.

If the editor knows scientist $i$, she has some prior information on the average quality of scientist $i$'s work. This is reflected in the model by assuming that the editor knows the value of $\mu_i$. For scientists she does not know, the editor is uncertain about the average quality of their work. All she knows is the distribution of average quality in the larger scientific community, which I also assume to be normal: $\mu_i \sim N(\mu, \sigma_{sc}^2)$.

Note that I assume the scientific community to be homogeneous: the scientific community is split in two groups (those known by the editor and those not known by the editor) but average paper quality follows the same distribution in both groups. If I assumed instead that scientists known by the editor write better papers on average the results would be qualitatively similar to those I present below. If scientists known by the editor write worse papers on average this would affect my results. However, since most journal editors are relatively central figures in their field (Crane 1967), this would be an implausible assumption except perhaps in isolated cases.

The editor's prior beliefs about the quality of a paper submitted by some scientist $i$ reflects this difference in information. If she knows the scientist she knows the value of $\mu_i$, and so her prior is $\pi(q_i \mid \mu_i) \sim N(\mu_i, \sigma_{qu}^2)$. If the editor does not know scientist $i$ she only knows the distribution of $\mu_i$, rather than its exact value. Integrating out the uncertainty over $\mu_i$ yields a prior $\pi(q_i) \sim N(\mu, \sigma_{qu}^2 + \sigma_{sc}^2)$ for the quality of scientist $i$'s paper.

When the editor receives a paper she sends it out for review. In the context of this model, the main purpose of the reviewer's report is to provide an estimate of the quality of the paper. But, I assume, even after reading the paper its quality cannot be established with certainty. Thus the reviewer's estimate $r_i$ of the quality $q_i$ is again a random variable. I assume that the reviewer's report is unbiased, i.e., its mean is the actual quality $q_i$ of the paper. Once again I use a normal distribution to reflect the uncertainty: $r_i \mid q_i \sim N(q_i, \sigma_{rv}^2)$.[5]

The editor uses the information from the reviewer's report to update her beliefs about the quality of scientist $i$'s paper. I assume that she does this by Bayes conditioning. Thus, her posterior beliefs about the quality of the paper are $\pi(q_i \mid r_i)$ if she does not know the author, and $\pi(q_i \mid r_i, \mu_i)$ if she does.

The posterior distributions are themselves normal distributions whose mean is a weighted average of $r_i$ and the prior mean, as given in proposition 4.1 (for a proof, see DeGroot 2004, section 9.5, or any other textbook that covers Bayesian statistics).

[5]The reviewer's report could reflect the opinion of a single reviewer, or the averaged opinion of multiple reviewers. The editor could even act as a reviewer herself, in which case the report reflects her findings which she has to incorporate in her overall beliefs about the quality of the paper. The assumption I make in the text can be used to cover any of these scenarios, as long as a given journal is fairly consistent in the number of reviewers used. If the number of reviewers is frequently different for different papers (and in particular when this difference correlates with the existence or absence of a connection between editor and author) the assumption of a fixed variance in the reviewer's report is unrealistic because a report from multiple reviewers may be thought to give more accurate information (reducing the variance) than a report from a single reviewer. Similarly, editors may use different reviewers in different roles (e.g., one reviewer to assess technical aspects of the paper and one reviewer to assess non-technical aspects). This is no problem for my model as long as the editor aggregates and converts these assessments in such a way that they can be represented together as a single assessment of the quality of the paper.

**Proposition 4.1.**

$$\pi(q_i \mid r_i) \sim N\left(\mu_i^U, \frac{(\sigma_{qu}^2 + \sigma_{sc}^2)\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}\right),$$
$$\pi(q_i \mid r_i, \mu_i) \sim N\left(\mu_i^K, \frac{\sigma_{qu}^2\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}\right),$$

*where*

$$\mu_i^U = \frac{\sigma_{qu}^2 + \sigma_{sc}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2} r_i + \frac{\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}\mu,$$
$$\mu_i^K = \frac{\sigma_{qu}^2}{\sigma_{qu}^2 + \sigma_{rv}^2} r_i + \frac{\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}\mu_i.$$

When does the editor choose to publish a paper? Here I assume that she publishes any paper whose posterior mean is above some threshold $q^*$. So a paper written by a scientist unknown to the editor is published if $\mu_i^U > q^*$ and a paper written by a scientist known to the editor is published if $\mu_i^K > q^*$. This corresponds to being at least 50% confident that the paper's quality is above the threshold. Other standards could be used (risk-averse standards might require more than 50% confidence that the paper is above some threshold, while risk-loving standards might require less; in these cases the threshold value needs to be adapted to keep the total number of accepted papers constant) but for my purposes here it does not much matter.

Now compare the probability that the paper of an arbitrary scientist $i$ unknown to the editor is published to the probability that the paper of an arbitrary scientist known by the editor is published. For this purpose it is useful to determine the probability distribution of the posterior means (see appendix D for proofs of this and subsequent results).

**Proposition 4.2.** *The posterior means are normally distributed, with* $\mu_i^U \sim N(\mu, \sigma_U^2)$ *and* $\mu_i^K \sim N(\mu, \sigma_K^2)$. *Here,*

$$\sigma_U^2 = \frac{(\sigma_{qu}^2 + \sigma_{sc}^2)^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2} \qquad \textit{and} \qquad \sigma_K^2 = \frac{\sigma_{qu}^4 + \sigma_{sc}^2(\sigma_{qu}^2 + \sigma_{rv}^2)}{\sigma_{qu}^2 + \sigma_{rv}^2}.$$

*Moreover, if* $\sigma_{sc}^2 > 0$ *and* $\sigma_{rv}^2 > 0$*, then* $\sigma_U^2 < \sigma_K^2$*.*

The main result of this section, which establishes the existence of connection bias in the model, is a consequence of proposition 4.2. It says that the editor is more likely to publish a paper written by an arbitrary author she knows than a paper written by an arbitrary author she does not know, whenever $q^* > \mu$ (for any positive value of $\sigma_{sc}^2$ and $\sigma_{rv}^2$). Since $q^* = \mu$ would mean that exactly half of all papers gets published, the condition amounts to a requirement that the journal's acceptance rate is less than 50%. This is true of most reputable journals in most fields (physics being a notable exception). When acceptance rates are above 50% editorial favoritism is also much less of a concern in the first place.

**Theorem 4.3.** *If* $q^* > \mu$*,* $\sigma_{sc}^2 > 0$*, and* $\sigma_{rv}^2 > 0$*, the acceptance probability for authors known to the editor is higher than the acceptance probability for authors unknown to the editor, i.e.,* $\Pr\left(\mu_i^K > q^*\right) > \Pr\left(\mu_i^U > q^*\right)$.

Theorem 4.3 shows that in the model I presented, any journal with an acceptance rate lower than 50% will be seen to display connection bias. Thus I have established the surprising result that an editor who cares only about the quality of the papers she publishes may end up publishing more papers by her friends and colleagues than by scientists unknown to her, even if her friends and colleagues are not, as a group, better scientists than average.

Why does this surprising result hold? The theorem follows immediately from proposition 4.2, which says that the distribution of $\mu_i^U$ is less "spread out" than the distribution of $\mu_i^K$ ($\sigma_U^2 < \sigma_K^2$). This happens because $\mu_i^U$ is a weighted average of $\mu$ and $r_i$, keeping it relatively close to the overall mean $\mu$ compared to $\mu_i^K$, which is a weighted average of $\mu_i$ and $r_i$ (which tend to differ from $\mu$ in the same direction).

Because the editor treats papers by authors she knows differently from papers by authors she does not know, authors unknown to the editor are arguably harmed. I pick up this point in section 4.3 and argue that this constitutes an epistemic injustice against those authors.

What I have shown so far is that an editor who uses information about the average quality of papers produced by scientists she knows in her acceptance decisions will find that scientists she knows produce on average more papers that meet her quality threshold. This is a subjective statement: the editor believes that more papers by scientists she knows meet her threshold. Does this translate into an objective effect? That is, does the extra information the editor has available about scientists she knows allow her to publish better papers from them than from scientists she does not know?

In order to answer this question I need to compare the average quality of accepted papers. More formally, I want to compare the expected value of the quality of a paper, conditional on meeting the publication threshold, given that the author is either known to the editor or not.

**Proposition 4.4.** *If $\sigma^2_{sc} > 0$, and $\sigma^2_{rv} > 0$, the average quality of accepted papers from authors known to the editor is higher than the average quality of accepted papers from authors unknown to the editor, i.e.,* $\mathbb{E}[q_i \mid \mu_i^K > q^*] > \mathbb{E}[q_i \mid \mu_i^U > q^*]$.

Proposition 4.4 shows that the editor can use the extra information she has about scientists she knows to improve the average quality of the papers published in her journal. In other words, the surprising result is that the editor's connection bias actually benefits rather than harms the readers of the journal. It is thus fair to say that, in the model, the editor can use her connections to "identify and capture high-quality papers", as Laband and Piette (1994a) suggest.[6]

[6]This result applies to connection bias only. Below I consider other biases the editor might have, which yields more nuanced conclusions.

To what extent does this show that the connection bias observed in reality is the result of editors capturing high-quality papers, as opposed to editors using their position of power to help their friends? At this point the model is seen to yield an empirical prediction. If connection bias is (primarily) due to capturing high-quality papers, the quality of papers by authors the editor knows should be higher than average, as shown in the model. If, on the other hand, connection bias is (primarily) a result of the editor accepting for publication papers written by authors she knows even though they do not meet the quality standards of the journal, then the quality of papers by authors the editor knows should (presumably) be lower than average.

If subsequent citations are a good indication of the quality of a paper,[7] a simple regression can test whether accepted papers written by authors with an author-editor connection have a higher or a lower average quality than papers without such a connection. This empirical test has been carried out a number of times, and the results univocally favor the hypothesis that editors use their connections to improve the quality of published papers (Laband and Piette 1994a, Smith and Dombrowski 1998, Medoff 2003).

Note that in the above results, nothing depends on the sizes of the variances $\sigma^2_{qu}$, $\sigma^2_{sc}$, and $\sigma^2_{rv}$. This is because these results are qualitative. The variances do matter when the acceptance rate and average quality of papers are compared quantitatively. For example, reducing $\sigma^2_{rv}$ (making the reviewer's report more accurate) makes the differences in the acceptance rate and average quality of papers smaller.

Note also that the results depend on the assumption that $\sigma^2_{sc}$ and $\sigma^2_{rv}$ are positive. What is the significance of these assumptions?

[7]Recall that I have remained neutral on how the notion of quality should be interpreted. If quality is simply defined as "the number of citations this paper would get if it were published" the connection between quality and citations is obvious. Even on other interpretations of quality, citations have frequently been viewed as a good proxy measure (Cole and Cole 1967, 1968, Medoff 2003). This practice has been defended by Cole and Cole (1971) and Clark (1957, chapter 3), and criticized by Lindsey (1989) and Heesen (forthcoming).

If $\sigma^2_{rv} = 0$, i.e., if there is no variance in the reviewer's report, the reviewer's report describes the quality of the paper with perfect accuracy. In this case the "extra information" the editor has about authors she knows is not needed, and so there is no difference in acceptance rate or average quality based on whether the editor knows the author. But it seems unrealistic to expect reviewer's reports to be this accurate (cf. the discussion of imperfect peer review in chapter 3).

If $\sigma^2_{sc} = 0$ there is either no difference in the average quality of papers produced by different authors, or learning the identity of the author does not tell the editor anything about the expected quality of that scientist's work. In this case there is no value to the editor (with regard to determining the quality of the submitted paper) in learning the identity of the author. So here also there is no difference in acceptance rate or average quality based on whether the editor knows the author.

Under what circumstances should the identity of the author be expected to tell the editor something useful about the quality of a submitted paper? This seems to be most obviously the case in the lab sciences. The identity of the author, and hence the lab at which the experiments were performed, can increase or decrease the editor's confidence that the experiments were performed correctly, including all the little checks and details that are impossible to describe in such a paper. In a scientific paper, "[a]s long as the conclusions depend at least in part on the results of some experiment, the reader must rely on the author's (and perhaps referee's) testimony that the author really performed the experiment exactly as claimed, and that it worked out as reported" (Easwaran 2009, p. 359).

But in other fields, in particular mathematics and some or all of the humanities, there is no need to rely on the author's reputation. This is because in these fields the paper itself is the contribution, so it is possible to judge papers in isolation of how or by whom they were created. Easwaran (2009) discusses this in detail for mathematics, and briefly (in his section 4)

for philosophy. And in fact there exists a norm that this is how they should be judged: "Papers will rely only on premises that the competent reader can be assumed to antecedently believe, and only make inferences that the competent reader would be expected to accept on her own consideration." (Easwaran 2009, p. 354).

Arguably then, the advantage (see theorem 4.3 and proposition 4.4) conferred by revealing identity information about the author to the editor applies only in certain fields. The relevant fields are those where part of the information in the paper is conferred on the authority of testimony, in particular those where experimental results are reported. Even in those fields, of course, what is being testified is supposed to be reproducible by the reader. But this is still different from the case in mathematics and the humanities, where a careful reading of a paper itself constitutes a reproduction of its argument. In these latter fields there is no relevant information to be learned from the identity of the author (i.e., $\sigma^2_{sc} = 0$), or, at least, the publishing norms in these fields suggest that their members believe this to be the case.

## 4.3 Bias As an Epistemic Injustice

The previous section discussed a formal model of editorial uncertainty about paper quality. The first main result, theorem 4.3, established the existence of connection bias in this model: authors known by the editor are more likely to see their paper accepted than authors unknown to the editor. The second main result, proposition 4.4, showed that connection bias benefits the readers of the journal by improving the average quality of accepted papers.

Despite the benefit to the readers, I claim that authors are harmed by connection bias. In this section I argue that an instance of connection bias constitutes an *epistemic injustice* in the sense of Fricker (2007). Then I argue that the editor is likely to display other biases as well, and that instances of these also constitute epistemic injustices.

The type of epistemic justice that is relevant here is *testimonial injustice*. Fricker (2007, pp. 17–23) defines a testimonial injustice as a case where a speaker suffers a credibility deficit for which the hearer is ethically and epistemically culpable, rather than being due to innocent error.

Testimonial injustices may arise in various ways. Fricker is particularly interested in what she calls "the central case of testimonial injustice" (Fricker 2007, p. 28). This kind of injustice results from a *negative identity-prejudicial stereotype*, which is defined as follows:

> A widely held disparaging association between a social group and one or more attributes, where this association embodies a generalization that displays some (typically, epistemically culpable) resistance to counter-evidence owing to an ethically bad affective investment. (Fricker 2007, p. 35)

Because the stereotype is widely held, it produces *systematic* testimonial injustice: the relevant social group will suffer a credibility deficit in many different social spheres.

Applying this to the phenomenon of connection bias, it is clear that this is not an instance of the central case of testimonial injustice. This would entail that there is some negative stereotype associated with scientists unknown to the editor, as a group, which is not normally the case. So I set the central case aside (I return to it below) and focus on the question whether connection bias can produce (non-central cases of) testimonial injustice.

Suppose scientist $i$ and scientist $i'$ tend to produce papers of the same quality, which is above average in the population ($\mu_i = \mu_{i'} > \mu$). Suppose further that the actual papers they have produced on this occasion are of the same quality ($q_i = q_{i'}$) and have received similar reviewer reports ($r_i = r_{i'}$). If scientist $i$ is not known to the editor, but scientist $i'$ is, then the paper written by scientist $i'$ is likely to be evaluated more highly by the editor.[8]

[8]The editor's posterior mean for the quality of scientist $i$'s paper is $\mu_i^U$ and her posterior

If the publication threshold $q^*$ is somewhere in between the two evaluations then only scientist $i'$ will have her paper accepted.

In this example, the scientists produced papers of equal quality that were evaluated differently. So scientist $i$ suffers a credibility deficit. This deficit is not due to innocent error, as it would be if, e.g., random variation led to different reviewer reports (i.e., $r_i < r_{i'}$). The deficit is also not due to the editor's use of generally reliable information about the two scientists, as it would be if there was a genuine difference in the average quality of the papers they produce (i.e., $\mu_i < \mu_{i'}$).

Is this credibility deficit suffered by scientist $i$ ethically and epistemically culpable on the part of the editor? On the one hand, as I stressed in section 4.2, the editor is simply making maximal use of the information available to her. It just so happens that she has more information about scientists she knows than about others. But that is hardly the editor's fault: she cannot be expected to know everyone's work. Is it incumbent upon her to get to know the work of every scientist who submits a paper?

This may well be too much to ask. But an alternative option is to remove all information about the authors of submitted papers. This can be done by using a triple-blind reviewing procedure, in which the editor does not know the identity of the author, and hence is prevented from using information about scientists she knows in her evaluation. Using such a procedure, at least all scientists are treated equally: any scientist who writes a paper of a given quality has the same chance of seeing that paper accepted.

So a credibility deficit occurs which harms scientist $i$: her paper is rejected. Moreover, it harms her specifically as an epistemic agent: the rejection of the paper reflects a judgment of the quality of her scientific work. And this harm could have been prevented by the editor by using a triple-blind reviewing procedure.

---

mean for scientist $i'$'s paper is $\mu^K_{i'} = \mu^K_i$, with $\mu^U_i < \mu^K_{i'}$ whenever $\sigma^2_{sc}(r_i - \mu_i) < (\sigma^2_{qu} + \sigma^2_{rv})(\mu_i - \mu)$. The claim in the text is then justified by the fact that $\Pr(\sigma^2_{sc}(r_i - \mu_i) < (\sigma^2_{qu} + \sigma^2_{rv})(\mu_i - \mu) \mid \mu_i > \mu) > 1/2$, assuming $\sigma^2_{sc} > 0$ and $\sigma^2_{rv} > 0$.

I conclude that the editor is ethically and epistemically culpable for this credibility deficit, and hence a testimonial injustice is committed against scientist $i$. However, one may insist that it cannot be the case that the editor is committing a wrong simply in virtue of using relevant information that is available to her. An evidentialist in particular may say that it cannot possibly be an epistemic wrong to take into account all relevant information.

I disagree, for the reasons just given, but I need not insist on this point. Even if it is granted that the editor does not commit an injustice by using the information that is available to her, the end result is still that scientist $i$ is harmed as an epistemic agent. She has produced a paper of equal quality to scientist $i'$'s, and yet it is not published.

Moreover, the presence of scientist $i'$ is irrelevant. Any time a paper from an author unknown to the editor is rejected which would have been accepted had the editor known the author (all else being equal), that author is harmed. So even if one insists that differential editorial treatment resulting from connection bias is not culpable on the part of the editor, connection bias still harms authors whenever it influences acceptance decisions.

A different objection points out that regardless of which reviewing procedure is used, a bias of some form occurs. If the editor uses information about the average quality of scientists, scientists with low average quality face a relatively high publication threshold, and so can complain about bias against them, whereas if the editor does not use such information, scientists with high average quality face a relatively high publication threshold, and so can complain about bias against them. In other words, scientists of high average quality prefer a double-blind reviewing procedure, whereas scientists of low average quality prefer a triple-blind reviewing procedure. But which one of these is more fair?[9] Here I judge intuitively that triple-blind review-

[9]Teddy Seidenfeld provided the following analogy to illustrate this difficulty. In court cases (at least in the United States), it is considered impermissible to use information about past criminal history to determine guilt or innocence. For doctors, it is considered irresponsible not to take into account information about a patient's medical history in

ing is fairer, but I leave providing a more detailed argument for this claim to future research.

In the model of section 4.2, and the above discussion, I assumed that connection bias is the only bias journal editors display. The literature on implicit bias suggests that this is not true. For example, "[i]f submissions are not anonymous to the editor, then the evidence suggests that women's work will probably be judged more negatively than men's work of the same quality" (Saul 2013, p. 45). Evidence for this claim is given by Wennerås and Wold (1997), Valian (1999, chapter 11), Steinpreis et al. (1999), Budden et al. (2008), and Moss-Racusin et al. (2012).[10] So women scientists are at a disadvantage simply because of their gender identity. Similar biases exist based on other irrelevant aspects of scientists' identity, such as race or sexual orientation (see Lee et al. 2013, for a critical survey of various biases in the peer review system). As Crandall (1982, p. 208) puts it: "The editorial process has tended to be run as an informal, old-boy network which has excluded minorities, women, younger researchers, and those from lower-prestige institutions".

I use *identity bias* to refer to these kinds of biases. Any time a paper is rejected because of identity bias (i.e., the paper would have been accepted if the relevant part of the author's identity had been different, all else being equal), a testimonial injustice occurs for the same reasons outlined above. Moreover, here the editor is culpable for having these biases.

Unlike instances resulting from connection bias, testimonial injustices resulting from identity bias can be instances of the central case of testimonial injustice, in which the credibility deficit results from a negative identity-prejudicial stereotype. The evidence suggests that negative identity-

determining future treatment. Why is reviewing academic papers more like a court and less like a doctor's office in this respect?

[10]These citations show that the work of women in academia is undervalued in various ways. None of them focus specifically on editor evaluations, but they support Saul's claim unless it is assumed that journal editors as a group are significantly less biased than other academics.

prejudicial stereotypes affect the way people (not just men) judge women's work, even when the person judging does not consciously believe in these stereotypes. Moreover, those who think highly of their ability to judge work objectively and/or are primed with objectivity are affected more rather than less (Uhlmann and Cohen 2007, Stewart and Payne 2008, p. 1333). Similar claims plausibly hold for biases based on race or sexual orientation. Biases based on academic affiliation are not usually due to negative identity-prejudicial stereotypes, as these do not generally affect other aspects of the scientist's life.

So both connection bias and identity bias are responsible for injustices against authors. This is one way to spell out the claim that authors are harmed when journal editors do not use a triple-blind reviewing procedure. This constitutes the first kind of argument for triple-blind reviewing which I mentioned in the introduction, and which I endorse based on these considerations.

## 4.4 The Effect of Bias on Quality

The second kind of argument I mentioned in the introduction claims that failing to use triple-blind reviewing harms the journal and its readers, because it would lower the average quality of accepted papers. In section 4.2 I argued that connection bias actually has the opposite effect: it increases average quality. In this section I complicate the model to include identity bias.

Recall that the editor displays identity bias if she is more or less likely to publish papers from a certain group of scientists based on some aspect of their identity, e.g., their gender. I incorporate this in the model by assuming the editor consistently undervalues members of one group (and overvalues the others). More precisely, she believes the average quality of papers produced by any scientist $i$ from the group she is biased against to be lower than it really is by some constant quantity $\varepsilon$. Conversely, the average quality of

papers written by any scientist not belonging to this group is raised by $\delta$.[11] So the editor has a different prior for the two groups; I use $\pi_A$ to denote her prior for the quality of papers written by scientists she is biased against, and $\pi_F$ for her prior for scientists she is biased in favor of.

As before, the editor may be familiar with a given scientist's work (i.e., she knows the average quality of that scientist's papers) or not. So there are now four groups. If scientist $i$ is known to the editor and belongs to the stigmatized group the editor's prior distribution on the quality of scientist $i$'s paper is $\pi_A(q_i \mid \mu_i) \sim N(\mu_i - \varepsilon, \sigma^2_{qu})$. If scientist $i$ is known to the editor but is not in the stigmatized group the prior is $\pi_F(q_i \mid \mu_i) \sim N(\mu_i + \delta, \sigma^2_{qu})$. If scientist $i$ is not known to the editor and is in the stigmatized group the prior is $\pi_A(q_i) \sim N(\mu - \varepsilon, \sigma^2_{qu} + \sigma^2_{sc})$. And if scientist $i$ is not known to the editor and not in the stigmatized group the prior is $\pi_F(q_i) \sim N(\mu + \delta, \sigma^2_{qu} + \sigma^2_{sc})$.[12]

The next few steps in the development are analogous to that in section 4.2. After the reviewer's report comes in the editor updates her beliefs about the quality of the paper, yielding the following posterior distributions.

---

[11] This is a simplifying assumption: one could imagine having biases against multiple groups of different strengths, or biases whose strength has some random variation, or biases which intersect in various ways (Collins and Chepp 2013, Bright et al. 2016). However, the assumption in the main text suffices to make the point I want to make. It should be fairly straightforward to extend my results to more complicated cases like the ones just described.

[12] Note that I assume that the editor displays bias against scientists in the stigmatized group regardless of whether she knows them or not. Under a reviewing procedure that is not triple-blind, the editor learns at least the name and affiliation of any scientist who submits a paper. This information is usually sufficient to determine with reasonable certainty the scientist's gender. So at least for gender bias it seems reasonable to expect the editor to display bias even against scientists she does not know. Conversely, because negative identity-prejudicial stereotypes can work unconsciously, it does not seem reasonable to expect that the editor can withhold her bias from scientists she knows.

**Proposition 4.5.**

$$\pi_A(q_i \mid r_i, \mu_i) \sim N\left(\mu_i^{KA}, \frac{\sigma_{qu}^2\sigma_{rv}^2}{\sigma_{qu}^2+\sigma_{rv}^2}\right),$$
$$\pi_F(q_i \mid r_i, \mu_i) \sim N\left(\mu_i^{KF}, \frac{\sigma_{qu}^2\sigma_{rv}^2}{\sigma_{qu}^2+\sigma_{rv}^2}\right),$$
$$\pi_A(q_i \mid r_i) \sim N\left(\mu_i^{UA}, \frac{(\sigma_{qu}^2+\sigma_{sc}^2)\sigma_{rv}^2}{\sigma_{qu}^2+\sigma_{sc}^2+\sigma_{rv}^2}\right),$$
$$\pi_F(q_i \mid r_i) \sim N\left(\mu_i^{UF}, \frac{(\sigma_{qu}^2+\sigma_{sc}^2)\sigma_{rv}^2}{\sigma_{qu}^2+\sigma_{sc}^2+\sigma_{rv}^2}\right),$$

*where*

$$\mu_i^{KA} = \mu_i^K - \frac{\varepsilon\cdot\sigma_{rv}^2}{\sigma_{qu}^2+\sigma_{rv}^2}, \qquad \mu_i^{KF} = \mu_i^K + \frac{\delta\cdot\sigma_{rv}^2}{\sigma_{qu}^2+\sigma_{rv}^2},$$
$$\mu_i^{UA} = \mu_i^U - \frac{\varepsilon\cdot\sigma_{rv}^2}{\sigma_{qu}^2+\sigma_{sc}^2+\sigma_{rv}^2}, \qquad \mu_i^{UF} = \mu_i^U + \frac{\delta\cdot\sigma_{rv}^2}{\sigma_{qu}^2+\sigma_{sc}^2+\sigma_{rv}^2}.$$

As before, the paper is published if the posterior mean ($\mu_i^{KA}$, $\mu_i^{KF}$, $\mu_i^{UA}$, or $\mu_i^{UF}$) exceeds the threshold $q^*$. The respective distributions of the posterior means determine how likely this is. These distributions are given in the next proposition.

**Proposition 4.6.** *The posterior means are normally distributed, with*

$$\mu_i^{KA} \sim N\left(\mu - \frac{\varepsilon\cdot\sigma_{rv}^2}{\sigma_{qu}^2+\sigma_{rv}^2}, \sigma_K^2\right),$$
$$\mu_i^{KF} \sim N\left(\mu + \frac{\delta\cdot\sigma_{rv}^2}{\sigma_{qu}^2+\sigma_{rv}^2}, \sigma_K^2\right),$$
$$\mu_i^{UA} \sim N\left(\mu - \frac{\varepsilon\cdot\sigma_{rv}^2}{\sigma_{qu}^2+\sigma_{sc}^2+\sigma_{rv}^2}, \sigma_U^2\right),$$
$$\mu_i^{UF} \sim N\left(\mu + \frac{\delta\cdot\sigma_{rv}^2}{\sigma_{qu}^2+\sigma_{sc}^2+\sigma_{rv}^2}, \sigma_U^2\right).$$

This yields the within-group acceptance rates and the unsurprising result that the editor is less likely to publish papers by scientists she is biased against.

**Theorem 4.7.** *If $\varepsilon > 0$, $\delta > 0$, and $\sigma_{rv}^2 > 0$, the acceptance probability for authors the editor is biased against is lower than the acceptance probability for authors the editor is biased in favor of (keeping fixed whether or not the editor knows the author). That is,*

$$\Pr\left(\mu_i^{KA} > q^*\right) < \Pr\left(\mu_i^{KF} > q^*\right) \text{ and } \Pr\left(\mu_i^{UA} > q^*\right) < \Pr\left(\mu_i^{UF} > q^*\right).$$

Theorem 4.7 establishes the existence of identity bias in the model: authors that are subject to a negative identity-prejudicial stereotype are less likely to see their paper accepted than authors who are not. As I argued in section 4.3, whenever a paper is rejected due to identity bias this constitutes a testimonial injustice against the author.

Now I turn my attention to the effect that identity bias has on the average quality of accepted papers. In the current version of the model there is both connection bias and identity bias. Connection bias has been shown to have a positive effect on average quality (see section 4.2). Whether the net effect of connection bias and identity bias is positive or negative depends on various parameters, as I illustrate below.

The benchmark for judging the average quality of accepted papers under a procedure subject to connection bias and identity bias is a *triple-blind reviewing procedure*. Under this triple-blind procedure, the editor's prior distribution for the quality of any submitted paper is $\pi(q_i) \sim N(\mu, \sigma_{qu}^2 + \sigma_{sc}^2)$, i.e., the prior I used in section 4.2 when the author was unknown to the editor. Hence, under this procedure, the posterior is $\pi(q_i \mid r_i)$, the posterior mean is $\mu_i^U \sim N(\mu, \sigma_U^2)$, the probability of acceptance is $\Pr(\mu_i^U > q^*)$ and the average quality of accepted papers is $\mathbb{E}[q_i \mid \mu_i^U > q^*]$. As a result, the editor is assumed to display neither connection bias nor identity bias.

In contrast, I refer to the reviewing procedure that is subject to connection bias and identity bias as the *non-blind procedure*. The overall probability that a paper is accepted under the non-blind procedure depends on the relative sizes of the four groups. I use $p_{KA}$ to denote the fraction of scientists known to the editor that she is biased against, $p_{KF}$ for the fraction known to the editor that she is biased in favor of, $p_{UA}$ for unknown scientists biased against, and $p_{UF}$ for unknown scientists biased in favor of. These fractions are nonnegative and sum to one.

Let $A_i$ denote the event that scientist $i$'s paper is accepted under the non-blind procedure. The overall probability of acceptance under this procedure is

$$\begin{aligned}\Pr(A_i) = p_{KA}\Pr\left(\mu_i^{KA} > q^*\right) + p_{KF}\Pr\left(\mu_i^{KF} > q^*\right) \\ + p_{UA}\Pr\left(\mu_i^{UA} > q^*\right) + p_{UF}\Pr\left(\mu_i^{UF} > q^*\right).\end{aligned}$$

The average quality of accepted papers can then be written as $\mathbb{E}[q_i \mid A_i]$. I want to compare $\mathbb{E}[q_i \mid A_i]$ to $\mathbb{E}[q_i \mid \mu_i^U > q^*]$, the average quality of accepted papers under a triple-blind procedure.[13]

In the remainder of this section I assume that the editor's biases are such that she believes the average quality of all submitted papers to be equal to $\mu$. In other words, her bias against the stigmatized group is canceled out on average by her bias in favor of those not in the stigmatized group, weighted by the relative sizes of those groups:

$$(p_{KA} + p_{UA})\,\varepsilon = (p_{KF} + p_{UF})\,\delta.$$

I use the above equation to fix the value of $\delta$, reducing the number of free

[13]Expressions for $\Pr(A_i)$ and $\mathbb{E}[q_i \mid A_i]$ using only the parameter values and standard functions are given in lemma D.4 in appendix D. These expressions are used to generate the numerical results below.

parameters by one. The equation amounts to a kind of commensurability requirement for the two procedures because it guarantees that the editor perceives the average quality of submitted papers to be the same regardless of whether or not a triple-blind procedure is used.

As far as I can tell there are no interesting general conditions on the parameter values that determine whether the non-blind procedure or the triple-blind procedure will lead to a higher average quality of accepted papers. The question I will explore now, using some numerical examples, is how biased the editor needs to be for the epistemic costs of her identity bias to outweigh the epistemic benefits resulting from connection bias.

In order to generate numerical data values have to be chosen for the parameters. First I set $\mu = 0$ and $q^* = 2$ (recall that $\mu$ is the average quality of all submitted papers and $q^*$ the threshold for accepting papers). Since quality is an interval scale in this model, these choices are arbitrary. For the variances $\sigma^2_{qu}$ (of the quality of individual papers), $\sigma^2_{sc}$ (of the average quality of authors), and $\sigma^2_{rv}$ (of the accuracy of the reviewer's report), I choose a "small" and a "large" value (1 and 4 respectively).

For the sizes of the four groups, I assume that there is no correlation between whether the editor knows an author and whether the editor has a bias against that author (so, e.g., the percentage of women among scientists the editor knows is equal to the percentage of women among scientists the editor does not know). I consider two cases for the editor's identity bias: either she is biased against half the set of authors (and so biased in favor of the other half) or the group she is biased against is a 30 % minority.[14] Similarly, I consider the case in which the editor knows half of all scientists submitting papers, and the case in which the editor knows 30 % of them.

[14]Bruner and O'Connor (forthcoming) note that certain dynamics in academic life can lead to identity bias against groups as a result of the mere fact that they are a minority. Here I consider both the case where the stigmatized group is a minority (and is possibly stigmatized as a result of being a minority, as Bruner and O'Connor suggest) and the case where it is not (and so presumably the negative identity-prejudicial stereotype has some other source).

As a result, there are 32 possible settings of the parameters ($2^3$ choices for the variances times $2^2$ choices for the group sizes). Whether the triple-blind procedure or the non-blind procedure is epistemically preferable depends on the value of $\varepsilon$ (and the value of $\delta$ determined thereby).

It follows from proposition 4.4 that when $\varepsilon = 0$ the non-blind procedure helps rather than harms the readers of the journal by increasing average quality relative to the triple-blind procedure. If $\varepsilon$ is positive but relatively small, this remains true, but when $\varepsilon$ is relatively big, the non-blind procedure harms the readers. This is because the average quality of published papers under the non-blind procedure decreases continuously as $\varepsilon$ increases (I do not prove this, but it is easily checked for the 32 cases I consider).

The interesting question, then, is where the turning point lies. How big does the editor's bias need to be in order for the negative effects of identity bias on quality to cancel out the positive effects of connection bias?

I determine the value of $\varepsilon$ for which the average quality of published papers under the non-blind procedure and the triple-blind procedure is the same for each of the 32 cases. But reporting these numbers directly does not seem particularly useful, as $\varepsilon$ is measured in "quality points" which do not have a clear interpretation outside of the model.

To give a more meaningful interpretation of these values of $\varepsilon$ as measuring "size of bias", I calculate the average rate of acceptance of papers from authors the editor is biased against and the average rate of acceptance of papers from authors the editor is biased in favor of.[15] The difference between these numbers gives an indication of the size of the editor's bias: it measures (in percentage points, abbreviated pp) how many more papers the

[15]These are calculated without regard for whether the editor knows the author or not. In particular, the rate of acceptance for authors the editor is biased against is

$$\frac{p_{KA}\Pr\left(\mu_i^{KA} > q^*\right) + p_{UA}\Pr\left(\mu_i^{UA} > q^*\right)}{p_{KA} + p_{UA}}, \text{ and } \frac{p_{KF}\Pr\left(\mu_i^{KF} > q^*\right) + p_{UF}\Pr\left(\mu_i^{UF} > q^*\right)}{p_{KF} + p_{UF}}$$

is the rate of acceptance for authors the editor is biased in favor of.

editor accepts from authors she is biased in favor of, compared to those she is biased against.

This difference is reported for the 32 cases in figure 4.1. To provide a sense of scale for these numbers, I plot them against the acceptance rate that the triple-blind procedure would have for those values of the parameters, i.e., $\Pr(\mu_i^U > q^*)$.

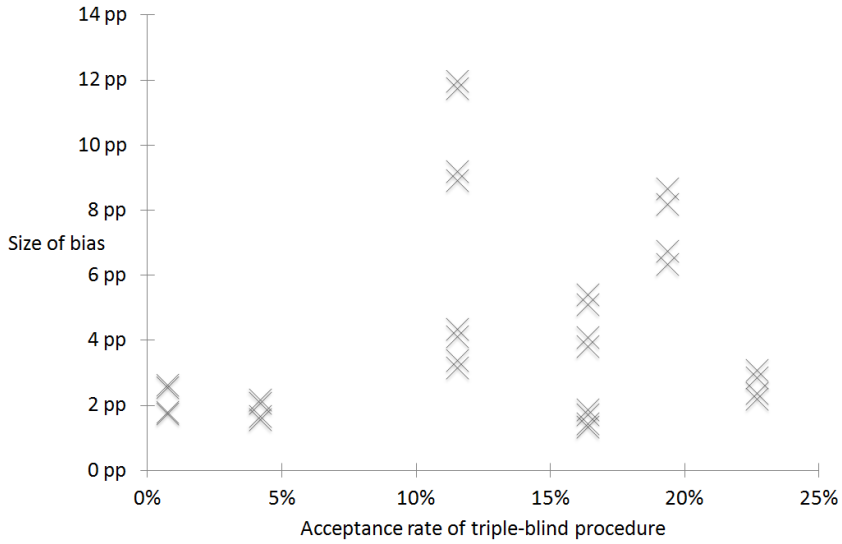


Figure 4.1: The minimum size of the editor's bias such that the quality costs of the non-blind procedure outweigh its benefits (given as a percentage point difference in acceptance rates), in 32 cases, plotted as a function of the acceptance rate of the corresponding triple-blind procedure.

Already with this small sample of 32 cases, a large variation of results can be observed. I illustrate this by looking at two cases in detail.

First, suppose that $\sigma_{qu}^2 = \sigma_{sc}^2 = 1$ and $\sigma_{rv}^2 = 4$. In this extreme case the triple-blind procedure has an acceptance rate as low as 0.72 %. If the groups are all of equal size ($p_{KA} = p_{KF} = p_{UA} = p_{UF} = 1/4$) then under the non-blind procedure the acceptance rate for authors the editor is biased in favor of needs to be as much as 2.66 pp higher than the acceptance rate for authors the editor is biased against, in order for the average quality under

the two procedures to be equal. Clearly a 2.66 pp bias is very large for a journal that only accepts less than 1 % of papers. If the bias is any less than that there is no harm to the readers in using the non-blind procedure.

Second, suppose that $\sigma^2_{qu} = \sigma^2_{sc} = 4$ and $\sigma^2_{rv} = 1$. Then the triple-blind procedure has an acceptance rate of 22.66 %. If, moreover, the editor knows relatively few authors ($p_{KA} = p_{KF} = 0.15$, $p_{UA} = p_{UF} = 0.35$) then the acceptance rate for authors the editor is biased in favor of needs to be only 2.23 pp higher than the acceptance rate for authors the editor is biased against, in order for the quality costs of the non-blind procedure to outweigh its benefits. For a journal accepting about 23 % of papers that means that even if the identity bias of the editor is relatively mild the journal's readers are harmed if the non-blind procedure is used.

Based on these results, and the fact that the parameter values are unlikely to be known in practice, it is unclear whether the non-blind procedure or the triple-blind procedure will lead to a higher average quality of published papers for any particular journal.[16] So in general it is not clear that an argument that the non-blind procedure harms the journal's readers can be made. At the same time, a general argument that the non-blind procedure helps the readers is not available either. Given this, I am inclined to recommend a triple-blind procedure for all journals because not doing so harms the authors.

If there was reason to believe that the editor's bias was very small, there might be a case for the non-blind procedure using considerations of average quality. Based on the empirical evidence I cited in section 4.3, it seems unlikely that any editor could make such a case convincingly today. But if identity bias were someday to be eliminated or severely mitigated, this question may be worth revisiting.

[16]Note that the evidence collected by Laband and Piette (1994a) does not help settle this question, as they do not directly compare the triple-blind and the non-blind procedure. Their evidence supports a positive epistemic effect of connection bias, but not a verdict on the overall epistemic effect of triple-blinding.

So far I have argued in this section that in the presence of the positive effect of connection bias on quality, the net effect of connection bias and identity bias on quality is unclear. But I argued in section 4.2 that the positive effect of connection bias may only exist in certain fields. In fields where papers rely partially on the author's testimony there is value in knowing the identity of the author. But in other fields such as mathematics and some of the humanities testimony is not taken to play a role—the paper itself constitutes the contribution to the field—and so arguably there is no value in knowing the identity of the author.

In those fields, then, there is no quality benefit from connection bias, but there is still a quality cost from identity bias. So here the strongest case for the triple-blind procedure emerges, as the non-blind procedure harms both authors and readers.

## 4.5 Conclusion

I have considered two types of arguments for triple-blind review.

I have argued that the non-blind procedure introduces differential treatment of scientific authors. In particular, editors are more likely to publish papers by authors they know (connection bias, theorem 4.3) and less likely to publish papers by authors they apply negative identity-prejudicial stereotypes to (identity bias, theorem 4.7). Whenever a paper is rejected as a result of one of these biases an epistemic injustice (in the sense of Fricker 2007) is committed against the author. This is an argument in favor of triple-blinding based on consequences for the author.

From the readers' perspective the story is more mixed. Generally speaking connection bias has a positive effect on the quality of published papers and identity bias a negative one. Thus whether the readers are better off under the triple-blind procedure depends on how exactly these effects trade off, which is highly context-dependent, or so I have argued. This yields a more

nuanced view than that suggested by either Laband and Piette (1994a), who focus only on connection bias, or by the argument for triple-blinding based on the consequences for the readers, which focuses only on identity bias.

However, in mathematics and some of the humanities there is arguably no positive quality effect from connection bias, as knowing about an author's other work is not taken to be relevant (Easwaran 2009). So here the negative effect of identity bias is the only relevant consideration from the readers' perspective. In this situation, considerations concerning the consequences for the author and considerations concerning the consequences for the readers point in the same direction: in favor of triple-blind review.

# Chapter 5

# Concluding Remarks

## 5.1 Journals, Priority, and Incentives

Journals play an important role as the brokers of scientific information. This dissertation has explored the incentive structure of science with regard to journal publications, both with regard to scientists-as-authors and scientists-as-editors.

This chapter reviews what we have learned. This section and the next consider the philosophical implications for the incentive structure of science and in particular the priority rule, which has a prominent place in the incentive structure. Section 5.3 draws some more methodological conclusions by looking at the way formal models have been used here. I finish the dissertation by briefly elaborating what policy makers might learn from it.

There is philosophical interest in exploring the epistemic consequences of aspects of scientific practice. In this case, a key aspect under investigation is the priority rule.

The *priority rule* is the feature of the social structure of science that determines the assignment of credit for scientific work. It says that the first scientist to make a particular contribution or discovery (prove a theorem, formulate a hypothesis, provide empirical support for a hypothesis, etc.)

takes the credit for it (Merton 1957, Strevens 2003).

In this dissertation I focused on the behavior of rational credit-maximizing scientists under the priority rule not because I think real scientists are credit maximizers. Rather my interest is in the incentive structure of science (cf. section 1.3). Through studying the behavior of rational credit-maximizing scientists I aim to reveal what kind of behavior the present incentive structure of science encourages, with an eye towards a normative appraisal of the incentive structure (not the behavior, which is rational by hypothesis).

The priority rule creates a pressure to publish (see sections 1.3 and 1.4). It is quite clear that this pressure has both good and bad epistemic consequences. Echoing the title of the dissertation and the title of chapter 3, we might say that the phrase "expediting the flow of knowledge" expresses the good consequences, while "rushing into print" captures the bad.

More formally speaking, I have shown in chapter 2 that the priority rule encourages the sharing of intermediate results, and can explain the so-called communist norm which mandates such sharing. This result may be viewed as "praise" for the priority rule, and as such coheres well with previous points of praise (Kitcher 1990, Dasgupta and David 1994, Strevens 2003, Boyer-Kassem and Imbert 2015).

In contrast, chapter 3 has illustrated one way in which the incentive structure created by the priority rule may fall short of being optimal. It systematically incentivizes scientists to work at a lower level of reliability than they would in an optimally functioning science. As a result, a relatively high number of erroneous results appears in the scientific literature. Moreover, I have argued that this downside is one that is unlikely to be preventable given a priority-based reward structure and imperfect peer review.

The pressure to publish also influences journal editors. As gatekeepers, their task is to determine which scientific work is to be published and what is to be rejected. They have an incentive to accept work of high impact and reject work that will either be ignored or found to be erroneous, as I assumed

in chapters 3 and 4.

In chapter 4 I argued that information about the identity of the author of a given paper may be useful to the editor in making this determination of likely impact. However, even a rational editor will display certain biases that harm the scientists who write scientific papers. I argue that on balance the incentive structure created by presenting the editor with identity information has bad consequences. This problem can be solved using triple-blind review, which, in a sense, protects the editor from herself.

## 5.2 Multiplying Priority

The priority rule played an essential role in chapters 2 and 3. In both of these chapters, I used rational choice models to investigate the question what rational credit-maximizing scientists would do in situations that are ubiquitous in science: you have done some work that you think is of scientific interest, but there is more to be done and you need to decide whether to (attempt to) publish now or wait. Chapter 3 focused on the question whether the work done so far has enough scientific merit to be published, whereas chapter 2 focused on the more strategic question whether, despite the work's merit, it would be worth keeping it secret in order to get a head start on a project that builds on the work already done.

In order to make claims about what rational credit-maximizers would do, I need to know how credit is distributed. The priority rule is helpful here, as it answers this question: the first scientist to make a particular contribution or discovery gets the credit for it.

The point I want to make in this section is that, despite its apparent specificity, the priority rule does not uniquely characterize the incentive structure of science. In saying that the first scientist to make a contribution takes the credit for it, at least two questions are left open: What counts as a contribution? And how much credit is given for a particular contribution?

My claim is that, properly speaking, *there is not one priority rule but multiple*. Only once it is specified what counts as a contribution and how much credit will be given for contributions has a determinate priority rule been given, and this can be done in multiple ways. Moreover, the details matter: different priority rules, and hence different reward structures, can incentivize different behavior, and may thus need to be appraised differently. This is already implicitly shown in this dissertation, as I argue next.

In chapter 2 I studied the question whether the priority rule gives scientists an incentive to share "intermediate results": scientific contributions that are part of a larger research project (such that scientists who do not share may have an advantage in completing subsequent parts of the project). While I answered this question in the affirmative, my answer explicitly relied on a particular specification of the priority rule. I assumed (first) that an intermediate result counts as a contribution and (second) that the amount of credit given for an intermediate result was equal for each part of the research project.

My conclusion depended on using this specification rather than some other one. If intermediate results do not count as contributions, there is no credit incentive to share anything at least until the entire research project is completed. If later intermediate results are rewarded with more credit than earlier ones, the incentive to share may also fail. So my claim that the priority rule incentivizes sharing, and that this reflects positively on it, is really a claim about a version of the priority rule that is simply not true of other versions.

In chapter 3 I argued that scientists have a credit incentive to "rush into print", that is to consistently publish scientific contributions that are less reliable than one would want them to be. Here also a claim is made about the incentive structure of science, and again it depends explicitly on the assumptions I made about what counts as a contribution and how much credit is given for a contribution.

In this case I assumed that both accurate and erroneous results can, in principle, count as contributions. I also assumed that accurate results are given more credit than erroneous results (at least on average). The latter assumption is actually not essential (if erroneous results were given more credit this would only exacerbate the problem of rushing into print) but the former is.

If erroneous results did not count as contributions (i.e., were not publishable) the problem of rushing into print as I analyze it would be avoided. However, I have argued that an incentive structure of this sort is in principle impossible, as it would require those who review the work to be able to consistently and perfectly predict whether a given result will later be disproven. So while changing this assumption would lead to a better incentive structure, doing so is impossible.

In section 3.4 I even considered a case in which scientists themselves have some limited control over their rewards, as they choose whether to aim for a risky high-impact contribution or a safer low-impact one. I showed, however, that there is a credit incentive to rush into print regardless of which type of work a given scientist has more affinity with.

So not only are there multiple priority rules, but there are subtle questions here about when it matters which version of the priority rule is used. This dissertation raises these questions but stops a long way short of settling them. Thus I propose for future research a more systematic investigation of the benefits and drawbacks of different versions of the priority rule. Such an investigation might fruitfully focus on the following three questions.

First, to what extent does existing work on the epistemic consequences of the priority rule depend on the version of the priority rule that is used? Existing work on the priority rule has identified various positive consequences of the priority rule. Perhaps most well-known is Strevens' work on scientists' choice of research programs. Other benefits have been suggested with regard to the speed of scientific discoveries (Dasgupta and David 1994), collabo-

ration (Boyer-Kassem and Imbert 2015), and sharing versus secrecy (Boyer 2014, Strevens forthcoming, and chapter 2 of this dissertation). A small number of potentially negative consequences of the priority rule have also been identified: the risk of herd behavior (Strevens 2013), the risk of increased fraud (Bright forthcoming), and the problem of rushing into print (chapter 3).

The work cited above does not include a systematic investigation of what is considered a contribution and how much credit is given for a contribution. For each of these papers it is to be expected that exploring these possibilities systematically will turn up cases where the identified consequence does not hold.

Second, insofar as epistemic consequences are specific to different versions of the priority rule, which of these consequences can occur simultaneously and which depend on incompatible versions? In other words, given a version of the priority rule, which of the above upsides and downsides actually apply? Can the benefits identified by Strevens, Dasgupta and David, Boyer-Kassem and Imbert, and myself all be realized simultaneously while avoiding the negative consequences, or (as seems likely) does any implementation involve some positive and some negative consequences?

Third, which version of the priority rule is best for science? Assuming that not all positive consequences can be realized simultaneously while avoiding all the negative consequences, answering this third question requires an overall evaluation of the different consequences. Such an evaluation may be difficult given the wide variety of types of consequences. However, the techniques used in this dissertation suggest a way to quantify these consequences, which makes them easier to compare. The next section takes up this and other methodological issues.

## 5.3 Formal Modeling As a Methodology

While the topics taken up in this dissertation are topics from the philosophy of science, the methodology that I have used is a combination of philosophical reasoning and mathematical modeling, drawing in particular on decision and game theory. While it is hardly surprising to find philosophical reasoning in a philosophy dissertation, mathematical models are still the exception rather than the rule. In this section I discuss some of the virtues of *formal methods*, emphasizing the ways I see this dissertation as exemplifying those virtues.

One virtue is that the construction of a formal model requires one to be precise about the phenomena one is describing (I am hardly the first one to point this out, see for instance Leitgeb 2013, section 7). It will not do to leave certain aspects of the problem one is describing underspecified, especially if one aims to *prove* results in one's model, as I have done throughout this dissertation.

In fact, some of the key results in this dissertation are stated as mathematical theorems. Theorems 2.1 and 2.2 show that there exists a credit incentive to share intermediate results in the model of chapter 2. Theorems 3.8 and 3.18 show that there exists a credit incentive to produce less reliable scientific work than is socially desirable in the model of chapter 3. And theorems 4.3 and 4.7 establish the existence of two kinds of biases in the way a journal editor chooses which papers to accept in the model of chapter 4, which I then go on to argue are instances of epistemic injustice.

It is only possible to prove that a particular fact obtains in a model if the model is stated with a level of specificity that makes that fact mathematically necessary. By requiring such a level of specificity the practice of mathematical modeling helps guard against equivocation. Both the fact that is proven to obtain in the model and the assumptions required to prove it are thereby illuminated.

From personal experience presenting and defending philosophical work that relies on models and theorems (from this dissertation and from previous

work), I know that the focus of discussion is almost always the assumptions that go into the model. And this seems entirely reasonable: if I present a model, a theorem, and a philosophical interpretation, and you disagree with my conclusion, given that the theorem is a mathematical necessity your only avenue for criticism is to argue that the assumptions that went into the model are inappropriate for the philosophical interpretation I have given.

Thus the discussion of these results naturally focuses on the assumptions, as it should. In this dissertation I have consequently spent significant time discussing the various assumptions I have made. Generally these fall into three categories.

First, the assumptions that I claim are empirically true or approximately true. For example, the assumption in chapter 2 that the productivity of a given scientist can be described as a Poisson process has been extensively tested and confirmed by Huber in a series of papers (Huber 1998a,b, 2001, Huber and Wagner-Döbler 2001a,b). My assumptions in chapter 3 about the relative credit and social value of erroneous scientific results are similarly supported by empirical work (Budd et al. 1998, Tatsioni et al. 2007). And with regard to the assumption made in chapters 3 and 4 that there is uncertainty about the quality or accuracy of a paper during peer review I have argued the opposite to require perfect prediction.

Second, assumptions which I argue (either formally or informally) do not affect the result. In chapter 2 I showed that whether or not scientists are assumed to know when other scientists have unpublished results is irrelevant to the rationality of their sharing behavior. In chapter 3 I showed that giving scientists a choice whether to go for high-impact results or low-impact results does not affect the problem of rushing into print.

Third, assumptions which are made for the sake of the argument. My assumptions about the motivations of the agents in my models fall in this category. In chapters 2 and 3 I assume that scientists are rational credit-maximizers, because my goal is to investigate what scientists have a credit

incentive to do (rather than what they actually do, given their complicated actual motivations, which is better studied empirically). And in chapter 4 I assume that the editor aims to publish high-quality papers, because my aims are to show that even editors with such pure motivations will display certain biases and to investigate the philosophical consequences of this fact.

Not all assumptions are defended in one of these ways, but most are. I hope to have given enough of a defense of the assumptions of each model that its conclusions will be taken seriously. My point here is that I agree that for any instance of formal methods in philosophy the assumptions should be carefully scrutinized, but that I disagree with those who feel that "all those unrealistics assumptions" are a weakness of formal methods as a philosophical methodology in general.

Everyone who makes a philosophical argument has to make assumptions. What I have tried to emphasize here is that one of the virtues of formal methods in philosophy is that it is much harder to attempt to obfuscate this fact. A mathematical theorem wears its assumptions on its sleeve in a way that philosophical arguments may not always do (although good ones usually do). Rather than criticizing the model's assumptions, we should think carefully about them. Where the assumptions apply, whatever theorems are true of the model apply as well, which may teach us something interesting. Where the assumptions fail to apply, we may ask what happens instead, and having stated the assumption may give us an interesting way to think about that.

A second virtue of formal methods in philosophy, and in particular of the kind of models I have considered in this dissertation, is that I can investigate counterfactuals. Experimenting with the social structure of science in real life is virtually impossible, but tweaking a parameter or changing an assumption can be done relatively easily in a formal model. In fact there is no reason to think that a model with "realistic" assumptions would be any easier or harder to analyze than a model which includes some interesting deviation

from reality.

A number of examples of this occur in chapter 3. I point out various ways to set the parameters of the model such that it is easy to show that there is *not* a credit incentive to rush into print. This happens, e.g., when peer review is perfect, or when the social value of erroneous results is so high that it is socially optimal to produce work with zero reliability. These scenarios are clearly unrealistic, but they are as easy (if not easier) to analyze in the model than more realistic scenarios.

While the above examples are not particularly practically useful, there are other instances in which the use of formal models to investigate counterfactuals can be used to evaluate policy proposals. The models in this dissertation yield a number of policy recommendations, which I discuss in section 5.4.

A third virtue of formal methods in philosophy is that they may raise *new philosophical questions*. An example of this occurs in section 5.2. There I point out that there is really not one priority rule, but multiple. This raises the questions what the epistemic consequences of different versions of the priority rule are, and which version is best overall.

These questions are raised as a direct result of my attempt to model the consequences of the priority rule for sharing behavior (chapter 2) and for rushing into print (chapter 3). It became clear that assuming "the priority rule applies" is not specific enough to prove what rational credit-maximizing scientists would do: further assumptions about what counts as a contribution and how much credit is given for a contribution are needed.

In this case, simply attempting to model the relevant phenomena revealed something that a more informal analysis might have missed. Now we know not only that there are multiple priority rules, but that which one is used makes a difference. Even if one does not agree with the assumptions I made in my models about the relevant version of the priority rule, now that these questions have been raised future work on the priority rule needs to take

them into account.

## 5.4 Policy Implications

Another way in which this dissertation differs from most others one might see in philosophy is in its immediate practical implications. This section draws out some of the *policy implications* of the dissertation. I should emphasize, however, that I interpret the word "policy" broadly: these are general recommendations about how the social organization of science might be improved. Some of these could conceivably be implemented by a single body (e.g., a national government, or a grant-giving agency), while others require coordinated action from working scientists, journal editors, governments, and agencies.

In chapter 2 I showed that, given some assumptions about how credit is distributed, scientists have a credit incentive to share their work, including "intermediate results" that are achieved along the way to completing some overarching research project. Recall that one of these assumptions stated that each intermediate result is worth an equal amount of credit.

What happens if this "equal credit" assumption is changed? If earlier stages are worth more credit than later stages there is still an incentive to share (cf. Banerjee et al. 2014, corollary 2.2). But if later stages are worth more than earlier ones there may be a credit incentive to keep intermediate results secret.

Hence, if it appears that scientists are not sharing as much as policy makers want, a clear recommendation is available: give more credit for earlier stages of research projects and less for later stages (relatively speaking). This will strengthen scientists' incentive to share.

Since "giving credit" is something a scientific community does (not an individual agent or body) this recommendation may be hard to implement. On the other hand, there are at least some bodies that have significant indi-

vidual influence over what kind of scientific work is rewarded (e.g., the NIH in medical science, the NSF in many other fields, and the various committees that award the Nobel Prizes). If one of these bodies started a campaign for more recognition for scientists who made preliminary contributions to important scientific results, such an initiative may well percolate through the rest of the reward structure of science.

In chapter 3 I showed that there is a credit incentive to sacrifice reliability in favor of speed and/or impact. In section 3.5 I argued in some detail that there might not be a solution that completely eliminates this misalignment in the incentive structure of science. However, there are certain things that can be done to mitigate it. Other things being equal, improvements in the peer review system (accepting more accurate results, rejecting more erroneous results) reduce the misalignment between credit incentives and the social optimum.

Alternatively or additionally, bringing the average credit given for erroneous results more in line with their social value also reduces the misalignment. Presumably this means lowering the average credit for erroneous results. But note that this has the ethically dubious implication of penalizing scientists for honest mistakes.

This last point echoes a theme from chapter 4. If we focus our efforts on optimizing the scientific content of journal pages (maximizing the social value of scientific research in chapter 3, or maximizing the average quality of published papers in chapter 4) we may thereby hurt scientific authors.

This suggests that at least in some cases there is a tradeoff between socio-epistemic and ethical consequences of science policy. In this dissertation, I have made no attempts to adjudicate this tradeoff. While I have tended to focus on the socio-epistemic side of these issues, here I want to explicitly disavow the implication that I think that socio-epistemic considerations should be privileged over ethical ones. For a more sustained treatment of how ethical issues interact with epistemic ones see, e.g., Fricker (2007), Mills (2007),

and subsequent work.

Little more needs to be said about the policy implications of chapter 4. I claim that journals that do not already do so should institute triple-blind reviewing. I also endorse the claim that biases of any kind should be removed wherever possible, and reduced as much as possible elsewhere, but I provide no new suggestions for doing so. If, somehow, editors could be trusted to be completely unbiased, a case could be made that triple-blind reviewing would no longer be necessary. But the evidence suggests that we are not anywhere near that point yet.

# Appendix A

# A Unique Nash Equilibrium

Let $n \geq 2$ be the number of scientists and $k \geq 1$ the number of stages. Let $G^p_{n,k}$ denote the game of perfect information and let $G^m_{n,k}$ denote the game of imperfect information, as described in sections 2.4 and 2.5.

As is commonly done in game theory, I use $u_i(s_i, s_{-i})$ to denote the payoff (expected units of credit at the end of the game) to scientist $i$ if $s_i$ gives her strategy and $s_{-i}$ gives the strategies of all scientists other than $i$ (call this an "incomplete strategy profile").

One strategy is of particular interest. Let $s^E_i$ denote the strategy for scientist $i$ in which she plays $E$ (that is, shares and claims credit for her most recently completed stage) at every decision node in $G^p_{n,k}$ or at every information set in $G^m_{n,k}$.[1] Let $s^E_{-i}$ denote the incomplete strategy profile (in either game) where every scientist $i'$ other than scientist $i$ plays strategy $s^E_{i'}$. Let $S^E$ denote the strategy profile (in either game) in which every scientist $i$ plays strategy $s^E_i$.

**Lemma A.1.** *In both $G^p_{n,k}$ and $G^m_{n,k}$, for any scientist $i$, the payoff when every scientist always shares any stages she completes immediately is*

[1] Technically, then, $s^E_i$ denotes two strategies: one for each game. But they share a lot of features which I use below.

$$u_i\left(s_i^E, s_{-i}^E\right) = k\frac{\lambda_i}{\lambda}.$$

*Proof.* Scientist $i$ is the first to complete stage 1 with probability $\lambda_i/\lambda$. If she does she immediately claims one unit of credit. If any other scientist completes stage 1 before scientist $i$, that scientist immediately claims one unit of credit. Thus scientist $i$'s expected credit from the first stage is $\lambda_i/\lambda$. Then all scientists simultaneously start working on the next stage. So by the same reasoning, scientist $i$'s expected credit from any given stage is $\lambda_i/\lambda$. The result follows. □

The next lemma shows that if not every scientist always shares, scientists who always share get a higher payoff than they do in lemma A.1.

**Lemma A.2.** *Let $s_{-i}$ denote any incomplete strategy profile such that at least one scientist $i'$ plays some strategy other than $s_{i'}^E$ (this can be either a different pure strategy, or any mixed strategy which plays strategy $s_{i'}^E$ with probability less than one). In the case of $G_{n,k}^p$, add the further assumption that this involves a deviation on the equilibrium path, i.e., there is at least one scientist $i'$ who plays a strategy $s_{i'}$ (or a mixed strategy in which $s_{i'}$ is played with positive probability) such that if every other scientist $i''$ plays strategy $s_{i''}^E$ then there is a positive probability of reaching a decision node at which strategy $s_{i'}$ plays strategy $H$. Then in both $G_{n,k}^p$ and $G_{n,k}^m$*

$$u_i\left(s_i^E, s_{-i}\right) > k\frac{\lambda_i}{\lambda}.$$

*Proof.* Note that in the case described by lemma A.1, i.e., when the strategy profile $S^E$ is being played, the outcome of a single instance of the game can be described by a sequence $(i_1, i_2, \ldots, i_k)$, where the first member denotes the first scientist who completes a stage, the second member the second scientist to complete a stage (not necessarily a different scientist than the first), and so on. Because every scientist $i$ plays strategy $s_i^E$, each member of

the sequence also denotes the claiming of one unit of credit by that scientist. The probability of such a sequence describing the outcome of the game is

$$\frac{\lambda_{i_1}}{\lambda} \cdot \frac{\lambda_{i_2}}{\lambda} \cdots \frac{\lambda_{i_k}}{\lambda}.$$

Now suppose that there is at least one scientist $i'$ playing a strategy different from $s_{i'}^E$. Let $s_{i'} \neq s_{i'}^E$ be some strategy that scientist $i'$ plays with some probability $p > 0$ (where $p = 1$ if scientist $i'$ plays a pure strategy), and assume that $s_{i'}$ involves a deviation on the equilibrium path in the case of $G_{n,k}^p$.

A sequence like $(i_1, i_2, \ldots, i_k)$ can still be used to describe the first $k$ scientists to complete a stage, but because not everyone always claims credit, this may not completely describe the outcome of the game: if a scientist completed a stage but did not claim credit for it either immediately or later, it is possible that not all $k$ units of credit have been claimed after $k$ scientists have completed a stage.

However, regardless of whether credit is being claimed, the probability of the sequence remains unchanged due to the memorylessness property of the exponential distribution. Moreover, because scientist $i$ plays strategy $s_i^E$, she is still claiming a unit of credit whenever she occurs in the sequence. Thus, all possible sequences $(i_1, i_2, \ldots, i_k)$ still occur with the same probability, and scientist $i$ claims the same amount of credit in them. So scientist $i$ now expects to accrue $k\lambda_i/\lambda$ units of credit during the time it takes for $k$ scientists to complete a stage.

But, by assumption, there is at least one sequence $(i_1, i_2, \ldots, i_k)$ in which $i'$ occurs and (with probability $p$) plays strategy $H$ at the corresponding decision node or information set, and $i'$ does not occur in the remainder of that sequence. As a result, at the end of that sequence at most $k-1$ units of credit have been claimed. In the remainder of that game, there is a positive probability (at least $\lambda_i/\lambda$, the probability that she is the very next one to complete a stage) that scientist $i$ gains more credit, credit that she would not

have obtained if scientist $i'$ had played strategy $s_{i'}^E$. Since $p > 0$ and $\lambda_{i''} > 0$ for all $i''$, it follows that

$$u_i\left(s_i^E, s_{-i}\right) \geq k\frac{\lambda_i}{\lambda} + \frac{\lambda_{i_1}}{\lambda} \cdot \frac{\lambda_{i_2}}{\lambda} \cdots \frac{\lambda_{i_k}}{\lambda} \cdot p \cdot \frac{\lambda_i}{\lambda} > k\frac{\lambda_i}{\lambda}. \qquad \square$$

**Theorem A.3.** *Let $S$ be any strategy profile for $G_{n,k}^m$ other than $S^E$, or let $S$ be any strategy profile for $G_{n,k}^p$ that involves deviations on the equilibrium path relative to $S^E$. Then there exists at least one scientist $i$ playing strategy $s_i \neq s_i^E$ such that she would be strictly better off playing strategy $s_i^E$:*

$$u_i\left(s_i^E, s_{-i}\right) > u_i\left(s_i, s_{-i}\right).$$

*Proof.* Note that the game is zero-sum: regardless of strategies, there are $k$ units of credit to be divided, and so if one scientist's payoff increases, another's decreases. Combined with lemmas A.1 and A.2 this yields the theorem. Distinguish three cases:

1. There is only one scientist $i$ playing a (pure or mixed) strategy $s_i \neq s_i^E$. Then every scientist $i'$ other than scientist $i$ is playing strategy $s_{i'}^E$ and so by by lemma A.2 is getting a payoff greater than $k\lambda_{i'}/\lambda$. Because the game is zero-sum, it follows that $u_i(s_i, s_{-i}) < k\lambda_i/\lambda$. By lemma A.1, $u_i(s_i^E, s_{-i}) = k\lambda_i/\lambda$, and the result follows.

2. There is at least one scientist $i'$ playing strategy $s_{i'}^E$ and at least two scientists playing some other strategy. Then any scientist $i'$ who is playing strategy $s_{i'}^E$ is getting a payoff greater than $k\lambda_{i'}/\lambda$ by lemma A.2. Because the game is zero-sum, at least one of the remaining scientists, say scientist $i$, must be getting a payoff less than $k\lambda_i/\lambda$. But if scientist $i$ changed her strategy to $s_i^E$, by lemma A.2 she would get a payoff greater than $k\lambda_i/\lambda$. So $u_i(s_i^E, s_{-i}) > k\lambda_i/\lambda > u_i(s_i, s_{-i})$.

3. Every scientist $i'$ is playing some strategy $s_{i'} \neq s_{i'}^E$. Because the game is zero-sum, it is impossible for every scientist $i'$ to be getting a greater

payoff than $k\lambda_{i'}/\lambda$. So there is at least one scientist, say scientist $i$, such that $u_i(s_i, s_{-i}) \leq k\lambda_i/\lambda$. By lemma A.2, $u_i(s_i^E, s_{-i}) > k\lambda_i/\lambda$, and the result follows. □

Theorem A.3 plays an important role in the proofs of the results in the main text.

*Proof of theorem 2.1.* Consider the game $G_{n,k}^p$. In any profile (of pure or mixed strategies) at least one scientist has an incentive to change her strategy, unless every scientist $i$ plays strategy $s_i^E$ or a strategy that deviates from $s_i^E$ only off the equilibrium path. Thus no profile is a Nash equilibrium unless every scientist $i$ plays strategy $s_i^E$ or a strategy that deviates from $s_i^E$ only off the equilibrium path. But since the backwards induction solution is a Nash equilibrium, it follows that in the backwards induction solution (which is guaranteed to exist for any finite game of perfect information) every scientist $i$ must play strategy $s_i^E$ or a strategy that deviates from $s_i^E$ only off the equilibrium path. So in the backwards induction solution every scientist immediately shares and claims credit for any stage she completes. (A direct proof that in the backwards induction solution every scientist plays strategy $E$ at every decision node—including those off the equilibrium path—is available from the author upon request.) □

*Proof of theorem 2.2.* Let $S$ be any profile (of pure or mixed strategies) for the game $G_{n,k}^m$. If $S \neq S^E$, then at least one scientist has an incentive to change her strategy, and so $S$ is not a Nash equilibrium.

That $S^E$ is a Nash equilibrium, and in fact a strict Nash equilibrium, also follows from theorem A.3 by considering the special case where $s_{-i} = s_{-i}^E$. This shows that a scientist $i$ who deviates unilaterally makes herself strictly worse off. □

*Proof of theorem 2.4.* It suffices to show that the game of imperfect information meets the criteria of Huttegger et al. (2014, theorem 4).

By theorem 2.2, the strategy profile in which every scientist plays strategy $E$ at every information set is the only strict Nash equilibrium of the game.

That the game is a weakly better reply game follows from theorem A.3. At any strategy profile, for at least one scientist $i$ whose strategy differs from $s_i^E$ switching to strategy $s_i^E$ is a better reply for her. This switch leads to a profile which is either the strict Nash equilibrium or in which the same is true for some other scientist. The result is a path of length at most $n$ from any profile to the strict Nash equilibrium, in which at each step along the path one scientist $i$ switches her strategy to $s_i^E$, and improves her payoff by doing so. □

# Appendix B

# Speed Versus Reliability

Define the following functions for all $p \in [0, 1]$:

$$\begin{aligned} C(p) &= c_a \beta p \lambda(p) + c_e \alpha (1-p) \lambda(p), \\ V(p) &= v_a \beta p \lambda(p) + v_e \alpha (1-p) \lambda(p), \\ c_{a,e}(p) &= ap\lambda(p) + e(1-p)\lambda(p). \end{aligned}$$

The function $C$ reflects the scientist's expected credit as a function of $p$: she aims to choose a value of $p$ that maximizes $C$. The function $V$ reflects the expected social value of the scientist's work. The family of functions $c_{a,e}$ is used for technical purposes.

The variable $p$ is the desired reliability, which can range from zero to one, and is chosen by the scientist. The parameters have the following interpretations: $c_a$ is the credit to the scientist for an accurate paper, $c_e$ is the credit for an erroneous paper, $v_a$ is the social value of an accurate paper, $v_e$ is the social value of an erroneous paper, $\beta$ is the probability that an accurate paper passes peer review, and $\alpha$ is the probability that an erroneous paper passes peer review.

The function $\lambda$ reflects the tradeoff between speed and reliability: $\lambda(p)$ is

the speed at which the scientist works given that the desired reliability is $p$. Here I state some sufficient conditions on $\lambda$ for the maxima of the functions $C$ and $V$ to be uniquely defined.

**Assumption B.1.** *$\lambda : [0,1] \to [0,\infty)$ is a function that takes nonnegative values. For all $p, p' \in [0,1]$, if $p < p'$, then $\lambda(p') < \lambda(p)$.*

It follows immediately that $\lambda$ is bounded (as $\lambda(p) \leq \lambda(0) < \infty$ for all $p \in [0,1]$).

**Assumption B.2.** *The function $\lambda$ is concave. This means that for every $p, p', t \in [0,1]$*

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

Assumption B.2 entails that $\lambda$ is continuous on $(0,1)$ (but not necessarily at the endpoints).

**Assumption B.3.** $\lim_{p\to 1} \lambda(p) = 0$.

This assumption asserts that the scientist cannot deliver perfect work (in the sense of zero probability of errors), no matter how slowly she works. She can, however, get arbitrarily close: due to assumption B.1, $\lambda(p) > 0$ for all $p < 1$.

**Lemma B.4.** *If assumptions B.1, B.2, and B.3 are satisfied, $\lambda$ is continuous on $[0,1]$.*

*Proof.* Due to assumption B.2, $\lambda$ is continuous on $(0,1)$. By assumption B.3, $\lim_{p\to 1} \lambda(p) = 0$. If $\lambda(1) > 0$, there must be some $p < 1$ such that $\lambda(p) < \lambda(1)$, contradicting assumption B.1. So $\lambda(1) = 0$ and $\lambda$ is continuous at $p = 1$.

It remains to show that $\lambda$ is continuous at $p = 0$. Because $\lambda$ is monotone and bounded, $\lim_{p\to 0} \lambda(p)$ exists. Because $\lambda$ is decreasing, $\lim_{p\to 0} \lambda(p) \leq$

$\lambda(0)$. Due to concavity, this inequality cannot be strict[1]. So $\lim_{p\to 0}\lambda(p) = \lambda(0)$, i.e., $\lambda$ is continuous at $p = 0$. □

**Lemma B.5.** *If assumptions B.1, B.2, and B.3 are satisfied, there exists $p_{a,e}$ such that $c_{a,e}(p_{a,e}) = \max_{p\in[0,1]} c_{a,e}(p)$. Moreover, if $a > 0$, then $p_{a,e} < 1$ uniquely maximizes $c_{a,e}$.*

*Proof.* By lemma B.4, $\lambda$ is continuous on $[0, 1]$. It follows that $c_{a,e}$ is continuous. By the extreme value theorem, $c_{a,e}$ attains its maximum, i.e., there exists $p_{a,e} \in [0, 1]$ such that $c_{a,e}(p_{a,e}) = \max_{p\in[0,1]} c_{a,e}(p)$.

Note that $a > 0$ implies that there is at least some value of $p$ for which $c_{a,e}(p) > 0$: if $e \geq 0$ this is true for all $p \in (0, 1)$ and if $e < 0$ this is true because

$$\frac{a-2e}{2(a-e)} \in (0,1), \text{ and}$$
$$c_{a,e}\left(\frac{a-2e}{2(a-e)}\right) = \frac{1}{2}a\lambda\left(\frac{a-2e}{2(a-e)}\right) > 0.$$

It follows that $c_{a,e}(p_{a,e}) > 0$. Since $\lambda(1) = 0$, $c_{a,e}(1) = 0 < c_{a,e}(p_{a,e})$. So $p_{a,e} \neq 1$.

To see that $a > 0$ implies uniqueness of the maximum, write $c_{a,e}$ as the product of two concave functions:

$$c_{a,e}(p) = (e + (a - e)p)\lambda(p),$$

where $\lambda$ is concave by assumption B.2 and $e + (a - e)p$ is concave because it is linear. The latter is nonnegative whenever $p \geq \frac{-e}{a-e}$ (unless $a = e$, but then it is positive for all $p$), where $\frac{-e}{a-e} < 1$. Note that assumption B.1 implies that $\lambda$ is not identically zero on $[\max\{0, \frac{-e}{a-e}\}, 1]$ and $a > 0$ implies that $e + (a-e)p$ is not identically zero on $[\max\{0, \frac{-e}{a-e}\}, 1]$. Moreover, on the interval

[1] If the limit is $\ell < \lambda(0)$, choose $p$ small enough so that $\lambda(p)$ is within $(\lambda(0) - \ell)/2$ of $\ell$. Then $\lambda(0)/2 + \lambda(p)/2 > \lambda(p/2)$, contradicting concavity.

$[\max\{0, \frac{-e}{a-e}\}, 1]$, the function $\lambda$ has a unique maximum at $\max\{0, \frac{-e}{a-e}\}$. So by Kantrowitz and Neumann (2005, theorem 4.iii), the function $c_{a,e}$ has a unique maximum on the interval $[\max\{0, \frac{-e}{a-e}\}, 1]$. Since $c_{a,e} < 0$ whenever $0 \leq p < \frac{-e}{a-e}$, it follows that $c_{a,e}$ has a unique maximum on $[0, 1]$. □

**Lemma B.6.** *If assumptions B.1, B.2, and B.3 are satisfied, $a > 0$ and $a > 2e$, then $p_{a,e} > 0$ (where $p_{a,e}$ uniquely maximizes $c_{a,e}$).*

*Proof.* Existence and uniqueness of $p_{a,e}$ follow from lemma B.5. It follows from $a > 2e$ that $\frac{a-2e}{2(a-e)} \in (0, 1)$. Because $\lambda$ is concave and $\lambda(1) = 0$,

$$\frac{a}{2(a-e)}\lambda(0) = \frac{a}{2(a-e)}\lambda(0) + \frac{a-2e}{2(a-e)}\lambda(1) \leq \lambda\left(\frac{a-2e}{2(a-e)}\right).$$

Hence,

$$c_{a,e}\left(\frac{a-2e}{2(a-e)}\right) = \frac{1}{2}a\lambda\left(\frac{a-2e}{2(a-e)}\right) \geq \frac{a^2}{4(a-e)}\lambda(0) > \frac{4e^2}{4e}\lambda(0) = c_{a,e}(0),$$

so $c_{a,e}$ is not maximized at $p = 0$. □

**Lemma B.7.** *If assumptions B.1, B.2, and B.3 are satisfied, and $a > 0$, the function $c_{a,0}$ (that is, $c_{a,e}$ with $e = 0$) is uniquely maximized at $p^* \in (0, 1)$, where the value of $p^*$ does not depend on the value of $a$, and $c_{a,0}$ is increasing on $[0, p^*]$ and decreasing on $[p^*, 1]$.*

*Proof.* The conditions of this lemma entail that the conditions of lemmas B.5, and B.6 are satisfied. Hence there exists $p^* \in (0, 1)$ that uniquely maximizes $c_{a,0}$. Because $c_{a,0}(p) = ap\lambda(p)$, $a$ is merely a scaling constant, so the maximum is unchanged when $a$ changes.

Note (as in the proof of lemma B.5) that $c_{a,0}$ is the product of two concave functions ($ap$ and $\lambda$) that are nonnegative and not identically zero on $[0, 1]$. So by Kantrowitz and Neumann (2005, theorem 4.ii), there exist $x_1$ and $x_2$

such that $c_{a,e}$ is increasing on $[0, x_1)$, constant on $(x_1, x_2)$, and decreasing on $(x_2, 1]$. Since the maximum is unique, it follows that $x_1 = x_2 = p^*$. □

**Lemma B.8.** *If assumptions B.1, B.2, and B.3 are satisfied, and $a > e \geq 0$, then $p_{a,e} \leq p^*$ (where $p_{a,e}$ uniquely maximizes $c_{a,e}$ and $p^*$ is as defined in lemma B.7).*

*Proof.* The existence and uniqueness of $p_{a,e}$ and $p^*$ follow from lemmas B.5 and B.7 respectively. Suppose for reductio that $p_{a,e} > p^*$. Then $p^*\lambda(p^*) > p_{a,e}\lambda(p_{a,e})$ by definition of $p^*$ and $e(1-p^*)\lambda(p^*) \geq e(1-p_{a,e})\lambda(p_{a,e})$ because $\lambda$ is decreasing and $e \geq 0$. So

$$\begin{aligned} c_{a,e}(p^*) &= ap^*\lambda(p^*) + e(1-p^*)\lambda(p^*) \\ &> ap_{a,e}\lambda(p_{a,e}) + e(1-p_{a,e})\lambda(p_{a,e}) = c_{a,e}(p_{a,e}), \end{aligned}$$

which contradicts the supposition that $p_{a,e}$ maximizes $c_{a,e}$. □

**Lemma B.9.** *If assumptions B.1, B.2, and B.3 are satisfied, and $a > 0 \geq e$, then $p_{a,e} \geq p^*$ (where $p_{a,e}$ uniquely maximizes $c_{a,e}$ and $p^*$ is as defined in lemma B.7).*

*Proof.* The existence and uniqueness of $p_{a,e}$ and $p^*$ follow from lemmas B.5 and B.7 respectively. Suppose for reductio that $p_{a,e} < p^*$. Then $p^*\lambda(p^*) > p_{a,e}\lambda(p_{a,e})$ by definition of $p^*$ and $e(1-p^*)\lambda(p^*) \geq e(1-p_{a,e})\lambda(p_{a,e})$ because $\lambda$ is decreasing and $e \leq 0$. So

$$\begin{aligned} c_{a,e}(p^*) &= ap^*\lambda(p^*) + e(1-p^*)\lambda(p^*) \\ &> ap_{a,e}\lambda(p_{a,e}) + e(1-p_{a,e})\lambda(p_{a,e}) = c_{a,e}(p_{a,e}), \end{aligned}$$

which contradicts the supposition that $p_{a,e}$ maximizes $c_{a,e}$. □

**Theorem B.10.** *If assumptions B.1, B.2, and B.3 are satisfied, and $c_a\beta > 0$ and $v_a\beta > 0$, then there exist unique values $p_C < 1$ and $p_V < 1$ that maximize the functions $C$ and $V$ respectively, i.e.,*

$$C(p_C) = \max_{p\in[0,1]} C(p) \qquad \textit{and} \qquad V(p_V) = \max_{p\in[0,1]} V(p).$$

*Proof.* Note that $C$ and $V$ are special cases of $c_{a,e}$, with $C = c_{c_a\beta,c_e\alpha}$ and $V = c_{v_a\beta,v_e\alpha}$. Because $c_a\beta > 0$ and $v_a\beta > 0$ the conditions of lemma B.5 apply to $C$ and $V$. The result follows immediately. □

**Theorem B.11.** *Let assumptions B.1, B.2, and B.3 be satisfied. Assume also that $c_e\alpha \geq v_e\alpha$ and either $v_a\beta \geq c_a\beta > c_e\alpha \geq 0$ or $c_a\beta \geq v_a\beta > 0 \geq c_e\alpha$. Define $p_C$ and $p_V$ as in theorem B.10. Then $p_C \leq p_V$.*

*Proof.* Both sets of conditions imply that $c_a\beta > 0$ and $v_a\beta > 0$, so theorem B.10 applies. Define $p_C$ and $p_V$ as in theorem B.10.

Consider first the case where $v_a\beta \geq c_a\beta > c_e\alpha \geq 0$. Because $c_a\beta > c_e\alpha \geq 0$, lemma B.8 applies to $C$, so $p_C \leq p^*$, where $p^*$ is as defined in lemma B.7. Suppose for reductio that $p_V < p_C$. Then $C(p_C) > C(p_V)$ by definition of $p_C$, $p_C\lambda(p_C) > p_V\lambda(p_V)$ by lemma B.7 (because $0 \leq p_V < p_C \leq p^*$), and $(1-p_C)\lambda(p_C) < (1-p_V)\lambda(p_V)$ by assumption B.1. Hence,

$$\begin{aligned} V(p_C) &= C(p_C) + (v_a\beta - c_a\beta)p_C\lambda(p_C) + (v_e\alpha - c_e\alpha)(1-p_C)\lambda(p_C) \\ &> C(p_V) + (v_a\beta - c_a\beta)p_V\lambda(p_V) + (v_e\alpha - c_e\alpha)(1-p_V)\lambda(p_V) \\ &= V(p_V), \end{aligned}$$

contradicting the supposition that $p_V$ maximizes $V$. So $p_C \leq p_V$.

Now consider the case where $c_a\beta \geq v_a\beta > 0 \geq c_e\alpha$. Because $v_a\beta > 0 \geq v_e\alpha$, lemma B.9 applies to $V$, so $p_V \geq p^*$. Suppose for reductio that $p_V < p_C$. Then $C(p_C) > C(p_V)$ by definition of $p_C$, $p_C\lambda(p_C) < p_V\lambda(p_V)$ by lemma B.7 (because $p^* \leq p_V < p_C \leq 1$), and $(1-p_C)\lambda(p_C) < (1-p_V)\lambda(p_V)$ by assumption B.1. Hence,

$$\begin{aligned} V(p_C) &= C(p_C) + (v_a\beta - c_a\beta)p_C\lambda(p_C) + (v_e\alpha - c_e\alpha)(1-p_C)\lambda(p_C) \\ &> C(p_V) + (v_a\beta - c_a\beta)p_V\lambda(p_V) + (v_e\alpha - c_e\alpha)(1-p_V)\lambda(p_V) \\ &= V(p_V), \end{aligned}$$

contradicting the supposition that $p_V$ maximizes $V$. So $p_C \leq p_V$. □

**Corollary B.12.** *Let assumptions B.1, B.2, and B.3 be satisfied. Assume also that $\beta > \alpha > 0$, $c_a = v_a > 0$, and $v_e < c_e \leq c_a$. Define $p_C$ and $p_V$ as in theorem B.10. Then $p_C \leq p_V$.*

*Proof.* If $c_e \geq 0$ then the assumptions of the corollary imply the first set of conditions of theorem B.11, whereas if $c_e < 0$ the assumptions of the corollary imply the second set of conditions of theorem B.11. □

A slightly stronger conclusion may be obtained with the following additional assumption.

**Assumption B.13** ($\lambda$ is differentiable)**.** *The function $\lambda$ is differentiable on* $(0, 1)$.

With this additional assumption the result from theorem B.11 can be strengthened to a strict inequality.

**Theorem B.14.** *Let assumptions B.1, B.2, B.3, and B.13 be satisfied. Assume also that $c_e\alpha > v_e\alpha$ and either $v_a\beta \geq c_a\beta > c_e\alpha \geq 0$ and $v_a\beta > 2v_e\alpha$, or $c_a\beta \geq v_a\beta > 0 \geq c_e\alpha$. Define $p_C$ and $p_V$ as in theorem B.10. Then $p_C < p_V$.*

*Proof.* Since the assumptions of theorem B.11 are satisfied, $p_C \leq p_V$. Both sets of assumptions imply that $v_a\beta > 2v_e\alpha$ and $v_a\beta > 0$, so by lemmas B.5 and B.6 $0 < p_V < 1$. If $p_C = 0$ this completes the proof, so assume that $p_C > 0$. Then $0 < p_C \leq p_V < 1$ so the maximum of $C$ is achieved in the

interior of its domain. This means that $C'(p_C) = 0$. To show that $p_C \neq p_V$ it suffices to show that $V'(p_C) \neq 0$.[2]

Consider first the case where $v_a\beta \geq c_a\beta > c_e\alpha \geq 0$ and $v_a\beta > 2v_e\alpha$. Because $c_a\beta > c_e\alpha \geq 0$, lemma B.8 applies to $C$, so $p_C \leq p^*$, where $p^*$ is as defined in lemma B.7. By assumption B.13, the derivative of $p\lambda(p)$ exists and is given by $p\lambda'(p)+\lambda(p)$. By lemma B.7, $p_C \leq p^*$ entails $p_C\lambda'(p_C)+\lambda(p_C) \geq 0$. By assumption B.1, $\lambda'(p_C) < 0$. Putting all of this together yields

$$\begin{aligned} V'(p_C) &= C'(p_C) + (v_a\beta - c_a\beta)(p_C\lambda'(p_C) + \lambda(p_C)) \\ &\quad + (v_e\alpha - c_e\alpha)(1-p_C)\lambda'(p_C) - (v_e\alpha - c_e\alpha)\lambda(p_C) \\ &\geq (v_e\alpha - c_e\alpha)(1-p_C)\lambda'(p_C) - (v_e\alpha - c_e\alpha)\lambda(p_C) > 0. \end{aligned}$$

Now consider the case where $c_a\beta \geq v_a\beta > 0 \geq c_e\alpha$. Because $c_a\beta > 0 \geq c_e\alpha$, lemma B.9 applies to $C$, so $p_C \geq p^*$. By lemma B.7, this entails $p_C\lambda'(p_C) + \lambda(p_C) \leq 0$. Hence

$$\begin{aligned} V'(p_C) &= C'(p_C) + (v_a\beta - c_a\beta)(p_C\lambda'(p_C) + \lambda(p_C)) \\ &\quad + (v_e\alpha - c_e\alpha)(1-p_C)\lambda'(p_C) - (v_e\alpha - c_e\alpha)\lambda(p_C) \\ &\geq (v_e\alpha - c_e\alpha)(1-p_C)\lambda'(p_C) - (v_e\alpha - c_e\alpha)\lambda(p_C) > 0. \end{aligned}$$ □

**Corollary B.15.** *Let assumptions B.1, B.2, B.3, and B.13 be satisfied. Assume also that $\beta > \alpha > 0$, $c_a = v_a > 0$, $v_e < c_e \leq c_a$, and $v_e \leq v_a/2$. Define $p_C$ and $p_V$ as in theorem B.10. Then $p_C < p_V$.*

*Proof.* If $c_e \geq 0$ then the assumptions of the corollary imply the first set of conditions of theorem B.14, whereas if $c_e < 0$ the assumptions of the corollary imply the second set of conditions of theorem B.14. □

[2]The derivatives $C'$ and $V'$ are guaranteed to exist for all $p \in (0,1)$ by assumption B.13 and the product rule.

# Appendix C

# Speed Versus Reliability and Impact

In this section I investigate a model in which there is a three-way tradeoff between speed, reliability, and impact. The scientist chooses the minimal acceptable reliability $p$, and the level of impact she wishes to achieve $c$ (equated with the amount of credit she will be given if she is successful), and her speed $\lambda$ is determined as a function of $p$ and $c$.

As before, $p$ is interpreted as a probability, so its domain is naturally constrained to the interval $[0, 1]$. The impact or credit $c$ is not similarly constrained. However, I assume that, at least for a given reliability $p$, there is a maximum impact that can be achieved. The following definitions formalize this setup.

**Definition C.1.** Let $\alpha, \beta \in [0, 1]$ and $c_e, v_e \leq 1$ be fixed parameters.

C.1.a. The *maximum impact function* is a function $\mu : [0, 1] \to [0, \infty)$.

C.1.b. The *domain (of admissible choices)* is the set $D = \{(p, c) \mid p \in [0, 1], c \in [0, \mu(p)]\}$.

C.1.c. The *speed function* is a function $\lambda : D \to [0, \infty)$.

C.1.d. The *credit function* is the function $C : D \to \mathbb{R}$ given by

$$C(p, c) = \beta pc\lambda(p, c) + \alpha c_e(1 - p)c\lambda(p, c)$$

for all $(p, c) \in D$.

C.1.e. The *(social) value function* is the function $V : D \to \mathbb{R}$ given by

$$V(p, c) = \beta pc\lambda(p, c) + \alpha v_e(1 - p)c\lambda(p, c)$$

for all $(p, c) \in D$.

I make a number of assumptions on the shape of $\lambda$. These assumptions are very similar to the ones I made before, although they have to be adapted to the new three-dimensional context.

**Assumption C.2.** *The function $\lambda$ is decreasing in both its arguments:*

*C.2.a. For all $p, p' \in [0, 1]$, if $p < p'$ and $c \leq \min\{\mu(p), \mu(p')\}$, then $\lambda(p', c) < \lambda(p, c)$.*

*C.2.b. For all $p \in [0, 1)$, if $c < c' \leq \mu(p)$, then $\lambda(p, c') < \lambda(p, c)$.*

Note that assumption C.2.b excludes the case where $p = 1$. This is because assumption C.4 below entails that $\lambda(1, c) = 0$ for all $c$, which is not decreasing if $\mu(1) > 0$.

**Lemma C.3.** *If assumption C.2 is satisfied,*

$$\lambda(p, c) \leq \lambda(p, 0) \leq \lambda(0, 0) < \infty,$$

*for any $(p, c) \in D$.*

*Proof.* The first inequality follows from assumption C.2.b and the second from assumption C.2.a. □

The next assumption has a role similar to assumption B.3. It requires that as the scientist gets close to perfect reliability ($p \to 1$), her speed vanishes ($\lambda \to 0$). This is required only when $c = 0$ (but see lemma C.9).

Additionally, the assumption requires (for fixed reliability) that as the scientist gets close to maximum impact ($c \to \mu(p)$), her speed vanishes ($\lambda \to 0$). This formalizes the intended interpretation of $\mu$ as the maximum impact that can be achieved at a given level of reliability.

**Assumption C.4.** *The function $\lambda$ vanishes as $p$ or $c$ approaches the edge of its domain $D$.*

*C.4.a.* $\lim_{p\to 1} \lambda(p, 0) = 0$.

*C.4.b. For all* $p \in [0, 1]$, $\lim_{c\to\mu(p)} \lambda(p, c) = 0$.

**Lemma C.5.** *If assumptions C.2.b and C.4.b are satisfied, $\lambda(p, \mu(p)) = 0$ for all $p \in [0, 1)$.*

*Proof.* Let $p \in [0, 1)$ and $\varepsilon > 0$. By assumption C.4.b, there exists a $c < \mu(p)$ such that $\lambda(p, c) < \varepsilon$. By assumption C.2.b and nonnegativity of $\lambda$, $0 \leq \lambda(p, \mu(p)) < \lambda(p, c) < \varepsilon$. So $\lambda(p, \mu(p)) = 0$. □

**Lemma C.6.** *If assumptions C.2 and C.4.b are satisfied, $\mu$ is decreasing on $[0, 1)$.*

*Proof.* Let $p < p' < 1$ and suppose for reductio that $\mu(p) \leq \mu(p')$. Note that it follows that $\mu(p) \leq \min\{\mu(p), \mu(p')\}$. By lemma C.5, $\lambda(p, \mu(p)) = \lambda(p', \mu(p')) = 0$. But by assumption C.2,

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

Contradiction. So $\mu(p') < \mu(p)$. □

**Lemma C.7.** *If assumptions C.2 and C.4.b are satisfied, $\mu$ is bounded from above on $[0, 1)$ by $\mu(0) < \infty$.*

*Proof.* Immediate from lemma C.6. □

**Lemma C.8.** *If assumptions C.2 and C.4.b are satisfied,* $\lambda(p,c) \leq \lambda(0,c) < \infty$ *for any* $(p,c) \in D$.

*Proof.* Let $(p,c) \in D$. By definition $c \leq \mu(p)$. By lemma C.7 $\mu(p) \leq \mu(0)$. So $c \leq \min\{\mu(p), \mu(0)\}$. So by assumption C.2.a $\lambda(p,c) \leq \lambda(0,c)$. □

**Lemma C.9.** *If assumptions C.2 and C.4 are satisfied,* $\lim_{p\to 1} \mu(p)$ *exists and* $\lim_{p\to 1} \lambda(p,c) = 0$ *for all* $c \in [0, \lim_{p\to 1} \mu(p)]$.

*Proof.* By lemma C.6, $\mu$ is decreasing, and by definition, $\mu$ is bounded from below (by 0). Hence $\lim_{p\to 1} \mu(p)$ exists.

Let $0 \leq c \leq \lim_{p\to 1} \mu(p)$ and $\varepsilon > 0$. By assumption C.4.a, there exists $p' < 1$ such that $\lambda(p,0) < \varepsilon$ for all $p \in (p',1)$. By assumption C.2.b and nonnegativity of $\lambda$, $0 \leq \lambda(p,c) \leq \lambda(p,0) < \varepsilon$. So $\lim_{p\to 1} \lambda(p,c) = 0$. □

**Lemma C.10.** *If assumptions C.2 and C.4 are satisfied,* $\lambda(1,c) = 0$ *for all* $c \leq \min\{\mu(1), \lim_{p\to 1} \mu(p)\}$.

*Proof.* Let $c \leq \min\{\mu(1), \lim_{p\to 1} \mu(p)\}$ and $\varepsilon > 0$. By lemma C.9, there exists $p < 1$ such that $\lambda(p,c) < \varepsilon$. By assumption C.2.a and nonnegativity of $\lambda$, $0 \leq \lambda(1,c) < \lambda(p,c) < \varepsilon$. So $\lambda(1,c) = 0$. □

**Assumption C.11.** *The function* $\lambda$ *is concave. More specifically, for any* $(p,c),(p',c') \in D$ *and* $t \in [0,1]$,

*C.11.a.* $(tp + (1-t)p', tc + (1-t)c') \in D$;

*C.11.b.* $t\lambda(p,c) + (1-t)\lambda(p',c') \leq \lambda(tp + (1-t)p', tc + (1-t)c')$.

It does not follow from the definition of the domain $D$ of $\lambda$ or the assumptions made so far that $(tp + (1-t)p', tc + (1-t)c') \in D$, but this is required for the idea of a concave function to make sense, hence assumption C.11.a. The next lemma characterizes the meaning of assumption C.11.a.

**Lemma C.12.** *The following are equivalent:*

1. *Assumption C.11.a;*
2. *The set $D$ is convex;*
3. *The function $\mu$ is concave.*

*Proof.* Note first that $tp + (1-t)p' \in [0,1]$ for all $p, p', t \in [0,1]$.

$1 \Rightarrow 2$: A set is convex if every convex combination of two points in the set is itself in the set. Assumption C.11.a requires exactly that.

$2 \Rightarrow 3$: Let $p, p', t \in [0,1]$. By definition, $(p, \mu(p)), (p', \mu(p')) \in D$. By 2, it follows that $(tp+(1-t)p', t\mu(p)+(1-t)\mu(p')) \in D$. So by definition of $D$, $t\mu(p) + (1-t)\mu(p') \leq \mu(tp + (1-t)p')$. So $\mu$ is concave.

$3 \Rightarrow 1$: Let $(p,c), (p',c') \in D$ and $t \in [0,1]$. Then

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

where the latter inequality follows from the concavity of $\mu$. So $(tp + (1-t)p', tc + (1-t)c') \in D$. □

**Corollary C.13.** *If assumptions C.2, C.4 and C.11.a are satisfied, $\mu(1) \leq \lim_{p\to 1} \mu(p)$, and hence lemmas C.6 and C.7 apply also when $p = 1$.*

*Proof.* By lemma C.9, $\lim_{p\to 1} \mu(p)$ exists. By lemma C.12, $\mu$ is concave. $\mu(1) \leq \lim_{p\to 1} \mu(p)$ follows from this (cf. footnote 1). □

**Lemma C.14.** *If assumptions C.2, C.4 and C.11 are satisfied, then for all $p \in [0,1]$ $\lim_{c\to 0} \lambda(p,c) = \lambda(p,0)$.*

*Proof.* Let $p \in [0,1]$. First consider the case $p = 1$. $\lambda(1,c) = 0$ for all $c \leq \mu(1)$ by lemma C.10, so $\lim_{c\to 0} \lambda(1,c) = 0 = \lambda(1,0)$.

If $p < 1$, $\lim_{c\to 0} \lambda(p,c)$ exists because $\lambda$ is decreasing in $c$ (assumption C.2.b) and bounded from above (lemma C.3). Lemma C.3 implies that $\lim_{c\to 0} \lambda(p,c) \leq \lambda(p,0)$. This inequality cannot be strict because $\lambda$ is concave (cf. footnote 1). □

**Lemma C.15.** *If assumptions C.2, C.4.b and C.11 are satisfied, then for all* $c \in [0, \mu(0))$ $\lim_{p\to 0} \lambda(p, c) = \lambda(0, c)$.

*Proof.* Let $c \in [0, \mu(0))$. $\lim_{p\to 0} \lambda(p, c)$ exists because $\lambda$ is decreasing in $p$ (assumption C.2.a) and bounded from above (lemma C.8). Lemma C.8 implies that $\lim_{p\to 0} \lambda(p, c) \leq \lambda(0, c)$. This inequality cannot be strict because $\lambda$ is concave (cf. footnote 1). □

**Lemma C.16.** *If assumptions C.2, C.4 and C.11 are satisfied,* $\lambda$ *is continuous.*

*Proof.* Because $\lambda$ is concave, it is continuous at any interior point of its domain. It remains to show that $\lambda$ is continuous on the borders, that is at those points $(p, c)$ with $p = 0$, $p = 1$, $c = 0$, or $c = \mu(p)$. I give a proof for the case $c = 0$ (the other cases are similar).

Note first that $\lambda$ is continuous when one of its arguments is held fixed: the function $\lambda(\cdot, 0)$ (the restriction of $\lambda$ along the $p$-axis) is continuous due to concavity (at least for non-extreme values of $p$), and the function $\lambda(p, \cdot)$ is continuous for any fixed value of $p$: for $0 < c < \mu(p)$ this follows from concavity, for $c = 0$ this follows from lemma C.14, and for $c = \mu(p)$ this follows from assumption C.4.b and lemma C.5.

Let $p \in (0, 1)$[1] and $\varepsilon > 0$. By the foregoing there exists $\delta_1 > 0$ such that $\lambda(p', 0) < \lambda(p, 0) + \varepsilon$ for every $p' \in (p - \delta_1, p + \delta_1)$. By assumption C.2.a, $\lambda(p', 0) < \lambda(p, 0) + \varepsilon$ for every $p' > p - \delta_1$. Similarly, there exists $\delta_2 > 0$ such that $\lambda(p, c') > \lambda(p, 0) - \varepsilon/2$ for every $c' < \delta_2$. And, given the particular value of $\delta_2$ just chosen, there exists $\delta_3 > 0$ such that $\lambda(p', \delta_2) > \lambda(p, \delta_2) - \varepsilon/2 > \lambda(p, 0) - \varepsilon$ for every $p' < p + \delta_3$.

Choose $\delta = \min\{\delta_1, \delta_2, \delta_3\}$. Let $(p', c') \in D$ be such that $0 < \|(p', c') - (p, 0)\| < \delta$. It follows that $p - \delta_1 < p' < p + \delta_3$ and $0 \leq c' < \delta_2$. Hence, using assumption C.2.b and the facts established in the previous paragraph,

[1] I set aside the cases where $p = 0$ or $p = 1$ to avoid having to worry about certain technicalities, but essentially the same proof works for those cases too.

$$\lambda(p,0) - \varepsilon < \lambda(p', \delta_2) < \lambda(p', c') \leq \lambda(p', 0) < \lambda(p,0) + \varepsilon.$$

So $|\lambda(p', c') - \lambda(p, 0)| < \varepsilon$. So $\lambda$ is continuous at $(p, 0)$. □

Recall from topology the notion of the *interior* of a set. The interior of a set $A$, written $\operatorname{int} A$ is the set of all points $x \in A$ such that $x$ is contained in an open subset of $A$. Any point $x \in A$ that is not in the interior of $A$ is called a *boundary point* of $A$. The interior of the domain $D$ is the set $\operatorname{int} D = \{(p, c) \mid p \in (0, 1), c \in (0, \mu(p))\}$.

**Lemma C.17.** *If assumptions C.2, C.4 and C.11 are satisfied, $\beta > 0$ and $\alpha c_e > 0$, there is a unique point $(p_C, c_C) \in D$ such that*

$$C(p_C, c_C) = \max_{(p,c) \in D} C(p, c).$$

*Moreover, either $(p_C, c_C) \in \operatorname{int} D$ or $p_C = 0$.*

*Proof.* Because $\lambda$ is continuous (by lemma C.16), $C$ is continuous as well. By the extreme value theorem, there exists a point $(p_C, c_C) \in D$ such that $C(p_C, c_C) = \max_{(p,c) \in D} C(p, c)$ (uniqueness will be shown below).

Note that $C(p, c) > 0$ for all $(p, c) \in \operatorname{int} D$. Conversely, $C(p, c) = 0$ if either $c = 0$, $c = \mu(p)$ (by lemma C.5), or $p = 1$ (by lemma C.10). Hence, either $(p_C, c_C) \in \operatorname{int} D$ or $p_C = 0$.

Let $(p', c') \neq (p_C, c_C)$ be any point in $D$. To show uniqueness of the maximum, it suffices to show that $(p', c')$ does not maximize $C$.

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

$$\begin{aligned} f(t) &= C\left(tp_C + (1-t)p', tc_C + (1-t)c'\right) \\ &= \left(\alpha c_e + (\beta - \alpha c_e)\left(tp_C + (1-t)p'\right)\right) \\ &\qquad \cdot \left(tc_C + (1-t)c'\right) \lambda\left(tp_C + (1-t)p', tc_C + (1-t)c'\right) \end{aligned}$$

for all $t \in [0,1]$. Because $C$ is maximized at $(p_C, c_C)$, $f$ is maximized at $t = 1$.

Note that $f$ can be written as the product of three concave and nonnegative functions: $\lambda$ is a concave function of $t$ as a consequence of assumption C.11, and $\alpha c_e + (\beta - \alpha c_e)(tp_C + (1-t)p')$ and $tc_C + (1-t)c'$ are linear functions of $t$ and hence also concave. Moreover, since either $p_C \neq p'$ or $c_C \neq c'$, at least one of the functions $\alpha c_e + (\beta - \alpha c_e)(tp_C + (1-t)p')$ and $tc_C + (1-t)c'$ has a unique maximum (e.g., if $c_C > c'$, $tc_C + (1-t)c'$ is maximized at $t = 1$). Finally, none of the three functions are identically zero on $[0,1]$. So it follows from Kantrowitz and Neumann (2005, theorem 4) that $f$ has a unique maximum at $t = 1$. Hence $C(p', c') = f(0) < f(1) = C(p_C, c_C)$. □

**Lemma C.18.** *If assumptions C.2, C.4 and C.11 are satisfied, $\beta > 0$ and $\alpha c_e \leq 0$, there is a unique point $(p_C, c_C) \in \operatorname{int} D$ such that*

$$C(p_C, c_C) = \max_{(p,c)\in D} C(p, c).$$

*Proof.* Because $\lambda$ is continuous (by lemma C.16), $C$ is continuous as well. By the extreme value theorem, there exists a point $(p_C, c_C) \in D$ such that $C(p_C, c_C) = \max_{(p,c)\in D} C(p, c)$ (uniqueness will be shown below).

Let $D_C^+ = \left\{(p,c) \in D \mid p \geq \frac{-\alpha c_e}{\beta - \alpha c_e}\right\}$. Because $\beta > 0$ and $\alpha c_e \leq 0$, $0 \leq \frac{-\alpha c_e}{\beta - \alpha c_e} < 1$. Hence $\operatorname{int} D_C^+ = \left\{(p,c) \mid \frac{-\alpha c_e}{\beta - \alpha c_e} < p < 1, 0 < c < \mu(p)\right\}$ is nonempty.

As the name suggests, the significance of $D_C^+$ is that it denotes the part of the domain where $C$ is nonnegative. More precisely, $C(p,c) > 0$ if $(p,c) \in \operatorname{int} D_C^+$, $C(p,c) = 0$ if $(p,c)$ is a boundary point of $D_C^+$, and $C(p,c) < 0$ if $(p,c) \in D \setminus D_C^+$. It follows that $(p_C, c_C) \in \operatorname{int} D_C^+$ and (since $\operatorname{int} D_C^+ \subseteq \operatorname{int} D$) that $(p_C, c_C) \in \operatorname{int} D$.

Let $(p', c') \neq (p_C, c_C)$ be any point in $D$. To show uniqueness of the maximum, it suffices to show that $(p', c')$ does not maximize $C$. If $(p', c') \notin \operatorname{int} D_C^+$ then the proof is done because $C(p', c') \leq 0 < C(p_C, c_C)$. So suppose $(p', c') \in \operatorname{int} D_C^+$.

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

$$\begin{aligned} f(t) &= C\left(tp_C + (1-t)p', tc_C + (1-t)c'\right) \\ &= \left(\alpha c_e + (\beta - \alpha c_e)\left(tp_C + (1-t)p'\right)\right) \\ &\quad \cdot \left(tc_C + (1-t)c'\right) \lambda\left(tp_C + (1-t)p', tc_C + (1-t)c'\right) \end{aligned}$$

for all $t \in [0,1]$. Because $C$ is maximized at $(p_C, c_C)$, $f$ is maximized at $t = 1$.

Note that $f$ can be written as the product of three concave and nonnegative functions: $\lambda$ is a concave function of $t$ as a consequence of assumption C.11, and $\alpha c_e + (\beta - \alpha c_e)(tp_C + (1-t)p')$ and $tc_C + (1-t)c'$ are linear functions of $t$ and hence also concave. Moreover, since either $p_C \neq p'$ or $c_C \neq c'$, at least one of the functions $\alpha c_e + (\beta - \alpha c_e)(tp_C + (1-t)p')$ and $tc_C + (1-t)c'$ has a unique maximum. Finally, none of the three functions are identically zero on $[0,1]$. So it follows from Kantrowitz and Neumann (2005, theorem 4) that $f$ has a unique maximum at $t = 1$. Hence $C(p', c') = f(0) < f(1) = C(p_C, c_C)$. □

**Theorem C.19.** *If assumptions C.2, C.4, and C.11 are satisfied and $\beta > 0$, then there exist unique points $(p_C, c_C)$ and $(p_V, c_V)$ that maximize the functions $C$ and $V$ respectively, i.e.,*

$$C(p_C, c_C) = \max_{(p,c)\in D} C(p,c) \qquad \textit{and} \qquad V(p_V, c_V) = \max_{(p,c)\in D} V(p,c).$$

*Moreover, either $(p_C, c_C) \in \operatorname{int} D$ or $p_C = 0$, and either $(p_V, c_V) \in \operatorname{int} D$ or $p_V = 0$.*

*Proof.* For the function $C$, the result follows immediately from lemma C.17 if $\alpha c_e > 0$ and from lemma C.18 if $\alpha c_e \leq 0$.

Since the function $V$ is identical to the function $C$ except that $c_e$ is

replaced with $v_e$, the result for $V$ follows from lemma C.17 if $\alpha v_e > 0$ and from lemma C.18 if $\alpha v_e \leq 0$. □

**Theorem C.20.** *Let assumptions C.2, C.4, and C.11 be satisfied. Assume also that $\beta > 0$ and $\beta \geq \alpha c_e \geq \alpha v_e$. Define $(p_C, c_C)$ and $(p_V, c_V)$ as in theorem C.19. Then either $p_C < p_V$ or $(p_C, c_C) = (p_V, c_V)$.*

*Proof.* Suppose for reductio that $(p_C, c_C)$ and $(p_V, c_V)$ are distinct (either $p_C \neq p_V$ or $c_C \neq c_V$) and $p_C \geq p_V$. Because they are distinct, and $(p_V, c_V)$ uniquely maximizes $V$,

$$\begin{aligned} V(p_V, c_V) &= (\alpha v_e + (\beta - \alpha v_e)p_V)c_V\lambda(p_V, c_V) \\ &> (\alpha v_e + (\beta - \alpha v_e)p_C)c_C\lambda(p_C, c_C) = V(p_C, c_C). \end{aligned}$$

Since $0 < \alpha v_e + (\beta - \alpha v_e)p_V \leq \alpha v_e + (\beta - \alpha v_e)p_C$ it follows that $c_V\lambda(p_V, c_V) > c_C\lambda(p_C, c_C)$. But then

$$\begin{aligned} C(p_V, c_V) &= V(p_V, c_V) + \alpha(c_e - v_e)(1 - p_V)c_V\lambda(p_V, c_V) \\ &> V(p_C, c_C) + \alpha(c_e - v_e)(1 - p_C)c_C\lambda(p_C, c_C) = C(p_C, c_C), \end{aligned}$$

contradicting the fact that $(p_C, c_C)$ maximizes $C$. So the supposition is false, which means that either $(p_C, c_C) = (p_V, c_V)$ or $p_C < p_V$. □

**Corollary C.21.** *Let assumptions C.2, C.4, and C.11 be satisfied. Assume also that $\beta > \alpha > 0$ and $v_e < c_e \leq 1$. Define $(p_C, c_C)$ and $(p_V, c_V)$ as in theorem C.19. Then $p_C \leq p_V$.*

*Proof.* The conditions of the corollary imply the conditions of theorem C.20. □

In order to rule out the case that $(p_C, c_C) = (p_V, c_V)$ (and thus conclude that $p_C < p_V$) an additional assumption is needed.

**Assumption C.22.** *The partial derivative of the function $\lambda$ with respect to its first argument exists on the interior of its domain, i.e., $\frac{\partial}{\partial p}\lambda(p,c)$ exists for all $(p,c) \in \operatorname{int} D$.*

Before the result can be proven, one more lemma is needed (which does not actually use assumption C.22).

**Lemma C.23.** *Let assumptions C.2, C.4, and C.11 be satisfied. Assume also that $\beta > 0$ and $\beta > 3\alpha v_e$. Define $(p_V, c_V)$ as in theorem C.19. Then $(p_V, c_V) \in \operatorname{int} D$.*

*Proof.* By theorem C.19 either $(p_V, c_V) \in \operatorname{int} D$ or $p_V = 0$. So it suffices to show that $p_V \neq 0$.

Suppose for reductio that $(0, c)$ maximizes $V$ for some $0 < c < \mu(0)$ and that $v_e > 0$. By assumption $\beta > 3\alpha v_e$ and so $t \in (0, \frac{1}{4})$, where $t$ is defined by

$$t = \frac{\beta - 3\alpha v_e}{4(\beta - \alpha v_e)} \qquad \text{and hence} \qquad 1 - t = \frac{3\beta - \alpha v_e}{4(\beta - \alpha v_e)}.$$

Using assumption C.11 and $\lambda(1,0) = 0$ (by lemma C.10),

$$(1-t)\lambda(0,c) = (1-t)\lambda(0,c) + t\lambda(1,0) \leq \lambda(t,(1-t)c).$$

But then

$$\begin{aligned} V(t,(1-t)c) &= \left(\beta t(1-t)c + \alpha v_e(1-t)^2\right)\lambda(t,(1-t)c) \\ &\geq (1-t)^2\left(\beta t c + \alpha v_e(1-t)c\right)\lambda(0,c) \\ &> \alpha v_e c \lambda(0,c) = V(0,c), \end{aligned}$$

where the second inequality follows because $t$ was chosen such that

$$(1-t)^2\left(\beta t + \alpha v_e(1-t)\right) - \alpha v_e = \frac{(\beta - 3\alpha v_e)^2(9\beta - 7\alpha v_e)}{64(\beta - \alpha v_e)^2} > 0.$$

Since $V(t,(1-t)c) > V(0,c)$, $V$ does not have a maximum at $(0,c)$. Contradiction. □

**Theorem C.24.** *Let assumptions C.2, C.4, C.11, and C.22 be satisfied. Assume also that $\beta > 0$, $\beta \geq \alpha c_e > \alpha v_e$ and $\beta > 3\alpha v_e$. Define $(p_C, c_C)$ and $(p_V, c_V)$ as in theorem C.19. Then $p_C < p_V$.*

*Proof.* Since all the conditions of theorem C.20 are satisfied, either $p_C < p_V$ or $(p_C, c_C) = (p_V, c_V)$. So it suffices to show that $(p_C, c_C) \neq (p_V, c_V)$.

Since the conditions of lemma C.23 are also satisfied, $(p_V, c_V) \in \operatorname{int} D$. If $(p_C, c_C) \notin \operatorname{int} D$ the proof is finished, so suppose $(p_C, c_C) \in \operatorname{int} D$.

Assumption C.22 then entails that the partial derivatives of $C$ and $V$ with respect to their first argument exist at $(p_C, c_C)$. Since $(p_C, c_C)$ is an extremum of $C$ achieved in the interior of its domain, $\frac{\partial}{\partial p} C(p_C, c_C) = 0$. By assumption C.2.a, $\frac{\partial}{\partial p}\lambda(p_C, c_C) < 0$. Hence

$$\begin{aligned}\frac{\partial}{\partial p} V(p_C, c_C) = \frac{\partial}{\partial p} C(p_C, c_C) &+ (\alpha c_e - \alpha v_e) c_C \lambda(p_C, c_C) \\ &+ (\alpha v_e - \alpha c_e)(1 - p_C) c_C \frac{\partial}{\partial p}\lambda(p_C, c_C) > 0.\end{aligned}$$

So $(p_C, c_C)$ does not maximize $V$. So $(p_C, c_C) \neq (p_V, c_V)$. □

**Corollary C.25.** *Let assumptions C.2, C.4, C.11, and C.22 be satisfied. Assume also that $\beta > \alpha > 0$, $v_e < c_e \leq 1$ and $v_e \leq 1/3$. Define $(p_C, c_C)$ and $(p_V, c_V)$ as in theorem C.19. Then $p_C < p_V$.*

*Proof.* The conditions of the corollary imply the conditions of theorem C.24. □

# Appendix D

# The Acceptance Probability and the Average Quality of Papers

**Proposition 4.2.** *$\mu_i^U \sim N(\mu, \sigma_U^2)$ and $\mu_i^K \sim N(\mu, \sigma_K^2)$. Moreover, $\sigma_U^2 < \sigma_K^2$ whenever $\sigma_{sc}^2 > 0$ and $\sigma_{rv}^2 > 0$.*

*Proof.* First consider the distribution of $r_i$. Since $r_i \mid q_i \sim N(q_i, \sigma_{rv}^2)$, $q_i \mid \mu_i \sim N(\mu_i, \sigma_{qu}^2)$, and $\mu_i \sim N(\mu, \sigma_{sc}^2)$, it follows that $r_i \mid \mu_i \sim N(\mu_i, \sigma_{qu}^2 + \sigma_{rv}^2)$ and $r_i \sim N(\mu, \sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2)$.

The latter can be used straightforwardly to determine the distribution of $\mu_i^U$. Since $r_i - \mu \sim N(0, \sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2)$ it follows that

$$\frac{\sigma_{qu}^2 + \sigma_{sc}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}(r_i - \mu) \sim N\left(0, \frac{(\sigma_{qu}^2 + \sigma_{sc}^2)^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}\right) \sim N\left(0, \sigma_U^2\right).$$

The result follows because $\mu$ is a constant and

$$\mu_i^U = \frac{\sigma_{qu}^2 + \sigma_{sc}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2} r_i + \frac{\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2} \mu = \frac{\sigma_{qu}^2 + \sigma_{sc}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2} (r_i - \mu) + \mu.$$

Determining the distribution of $\mu_i^K$ is slightly trickier because there are two random variables involved: $r_i$ and $\mu_i$. As noted above, $r_i \mid \mu_i \sim N(\mu_i, \sigma_{qu}^2 + \sigma_{rv}^2)$. Thus, writing $X_i = \frac{\sigma_{qu}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}(r_i - \mu_i)$,

$$X_i \mid \mu_i \sim N\left(0, \frac{\sigma_{qu}^4}{\sigma_{qu}^2 + \sigma_{rv}^2}\right).$$

Since

$$\mu_i^K = \frac{\sigma_{qu}^2}{\sigma_{qu}^2 + \sigma_{rv}^2} r_i + \frac{\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2} \mu_i = X_i + \mu_i$$

it remains to determine the convolution of $X_i$ and $\mu_i$. This can be done using the moment-generating function and the law of total expectation. Recall that the moment-generating function of an $N(m, s^2)$ distribution is given by $M(t) = \exp\{mt + \frac{1}{2}s^2t^2\}$. So the moment-generating function of $\mu_i^K$ is

$$\begin{aligned}
\mathbb{E}[\exp\{t\mu_i^K\}] &= \mathbb{E}[\exp\{t(X_i + \mu_i)\}] \\
&= \mathbb{E}[\mathbb{E}[\exp\{tX_i + t\mu_i\} \mid \mu_i]] \\
&= \mathbb{E}[\exp\{t\mu_i\}\mathbb{E}[\exp\{tX_i\} \mid \mu_i]] \\
&= \exp\left\{0t + \frac{1}{2}\frac{\sigma_{qu}^4}{\sigma_{qu}^2 + \sigma_{rv}^2}t^2\right\} \mathbb{E}[\exp\{t\mu_i\}] \\
&= \exp\left\{\frac{1}{2}\frac{\sigma_{qu}^4}{\sigma_{qu}^2 + \sigma_{rv}^2}t^2 + \mu t + \frac{1}{2}\sigma_{sc}^2 t^2\right\} \\
&= \exp\left\{\mu t + \frac{1}{2}\frac{\sigma_{qu}^4 + \sigma_{sc}^2(\sigma_{qu}^2 + \sigma_{rv}^2)}{\sigma_{qu}^2 + \sigma_{rv}^2}t^2\right\},
\end{aligned}$$

which is exactly the moment-generating function of the desired normal dis-

tribution.

Finally, note that

$$\sigma_U^2 = \frac{(\sigma_{qu}^2 + \sigma_{sc}^2)^2(\sigma_{qu}^2 + \sigma_{rv}^2)}{(\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2)(\sigma_{qu}^2 + \sigma_{rv}^2)},$$
$$\sigma_K^2 = \frac{(\sigma_{qu}^2 + \sigma_{sc}^2)^2(\sigma_{qu}^2 + \sigma_{rv}^2) + \sigma_{sc}^2\sigma_{rv}^4}{(\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2)(\sigma_{qu}^2 + \sigma_{rv}^2)}.$$

So $\sigma_U^2 < \sigma_K^2$ whenever $\sigma_{sc}^2 > 0$ and $\sigma_{rv}^2 > 0$ (and $\sigma_U^2 = \sigma_K^2$ otherwise, assuming the expressions are well-defined in that case). □

**Theorem 4.3.** $\Pr\left(\mu_i^K > q^*\right) > \Pr\left(\mu_i^U > q^*\right)$ *if* $q^* > \mu$, $\sigma_{sc}^2 > 0$, *and* $\sigma_{rv}^2 > 0$.

*Proof.* It follows from proposition 4.2 that

$$\Pr\left(\mu_i^K > q^*\right) = 1 - \Phi\left(\frac{q^* - \mu}{\sigma_K}\right) \text{ and } \Pr\left(\mu_i^U > q^*\right) = 1 - \Phi\left(\frac{q^* - \mu}{\sigma_U}\right),$$

where $\Phi$ is the distribution function (or cumulative density function) of a standard normal distribution. Since $\Phi$ is (strictly) increasing in its argument, and $\sigma_K > \sigma_U$ by proposition 4.2, the theorem follows immediately. □

In order to prove proposition 4.4 a number of intermediate results are needed.

**Lemma D.1.**

$$\mathbb{E}[q_i \mid \mu_i^U > q^*] = \mathbb{E}[\mu_i^U \mid \mu_i^U > q^*],$$
$$\mathbb{E}[q_i \mid \mu_i^K > q^*] = \mathbb{E}[\mu_i^K \mid \mu_i^K > q^*].$$

*Proof.* Because $\mu_i^U$ is simply an (invertible) transformation of $r_i$, it follows that

$$q_i \mid \mu_i^U \sim q_i \mid r_i \sim N\left(\mu_i^U, \frac{(\sigma_{qu}^2 + \sigma_{sc}^2)\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}\right).$$

The distribution of $q_i \mid \mu_i^K$ is a little trickier to find, because $\mu_i^K$ is a linear combination of two random variables, $r_i$ and $\mu_i$, and it is not obvious that learning $\mu_i^K$ is as informative as learning both $r_i$ and $\mu_i$. But using the known distributions of $q_i \mid \mu_i$ and $\mu_i^K \mid q_i, \mu_i$ and integrating out $\mu_i$ it can be shown that

$$q_i \mid \mu_i^K \sim q_i \mid r_i, \mu_i \sim N\left(\mu_i^K, \frac{\sigma_{qu}^2\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}\right).$$

The important point here is that $\mathbb{E}[q_i \mid \mu_i^x] = \mu_i^x$ both for $x = U$ and $x = K$.

Now the law of total expectation can be used to establish that

$$\mathbb{E}[q_i \mid \mu_i^x > q^*] = \mathbb{E}[\mathbb{E}[q_i \mid \mu_i^x] \mid \mu_i^x > q^*] = \mathbb{E}[\mu_i^x \mid \mu_i^x > q^*],$$

for $x = U, K$. □

Let $X \sim N(\mu, \sigma^2)$ be a normally distributed random variable. Then $X \mid X > a$ follows a *left-truncated normal distribution*, with left-truncation point $a$. As a result of lemma D.1 I am interested in the mean of left-truncated normal distributions. According to, e.g., Johnson et al. (1994, chapter 13, section 10.1), this mean can be expressed as

$$\mathbb{E}[X \mid X > a] = \mu + \sigma R\left(\frac{a - \mu}{\sigma}\right). \tag{D.1}$$

Here

$$R(x) = \frac{\phi(x)}{1 - \Phi(x)}$$

for all $x \in \mathbb{R}$, where $\phi$ is the probability density function of the standard normal distribution, and $\Phi$ is its distribution function. $R$ is the inverse of what is known in the literature (e.g., Gordon 1941) as *Mills' ratio*.

It follows from the definitions that $R(x) > 0$ for all $x \in \mathbb{R}$ and that

$$R'(x) = R(x)^2 - xR(x). \tag{D.2}$$

**Proposition D.2** (Gordon (1941))**.** *For all $x > 0$, $R(x) < \frac{x^2+1}{x}$.*

Proposition D.2 can be used to establish the next result.

**Proposition D.3.** *If $X \sim N(\mu, \sigma^2)$ and $Y \sim N(\mu, s^2)$ with $s > \sigma > 0$ then $\mathbb{E}[Y \mid Y > a] > \mathbb{E}[X \mid X > a]$.*

*Proof.* It suffices to show that the derivative $\frac{\partial}{\partial \sigma}\mathbb{E}[X \mid X > a]$ is positive for all $\sigma > 0$. Differentiating equation (D.1) (using equation (D.2)) yields

$$\frac{\partial}{\partial \sigma}\mathbb{E}[X \mid X > a] = \left(\left(\frac{a-\mu}{\sigma}\right)^2 + 1\right) R\left(\frac{a-\mu}{\sigma}\right) - \frac{a-\mu}{\sigma} R\left(\frac{a-\mu}{\sigma}\right)^2.$$

Since $R\left(\frac{a-\mu}{\sigma}\right) > 0$, $\frac{\partial}{\partial \sigma}\mathbb{E}[X \mid X > a] > 0$ if and only if

$$\left(\frac{a-\mu}{\sigma}\right)^2 + 1 - \frac{a-\mu}{\sigma} R\left(\frac{a-\mu}{\sigma}\right) > 0.$$

This is true whenever $\frac{a-\mu}{\sigma} \leq 0$ because then both terms in the sum are positive. Proposition D.2 guarantees that it is true whenever $\frac{a-\mu}{\sigma} > 0$ as well. □

**Proposition 4.4.** $\mathbb{E}[q_i \mid \mu_i^K > q^*] > \mathbb{E}[q_i \mid \mu_i^U > q^*]$ *whenever* $\sigma_{sc}^2 > 0$*, and* $\sigma_{rv}^2 > 0$.

*Proof.* By lemma D.1,

$$\mathbb{E}[q_i \mid \mu_i^U > q^*] = \mathbb{E}[\mu_i^U \mid \mu_i^U > q^*],$$
$$\mathbb{E}[q_i \mid \mu_i^K > q^*] = \mathbb{E}[\mu_i^K \mid \mu_i^K > q^*].$$

By proposition 4.2, $\mu_i^U \sim N(\mu, \sigma_U^2)$ and $\mu_i^K \sim N(\mu, \sigma_K^2)$, with $\sigma_U < \sigma_K$. Hence the conditions of proposition D.3 are satisfied, and the result follows. □

**Proposition 4.6.**

$$\begin{aligned}
\mu_i^{KA} &\sim N\left(\mu - \frac{\varepsilon \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}, \sigma_K^2\right), \\
\mu_i^{KF} &\sim N\left(\mu + \frac{\delta \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}, \sigma_K^2\right), \\
\mu_i^{UA} &\sim N\left(\mu - \frac{\varepsilon \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}, \sigma_U^2\right), \\
\mu_i^{UF} &\sim N\left(\mu + \frac{\delta \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}, \sigma_U^2\right).
\end{aligned}$$

*Proof.* Since $\mu_i^{KA}$ and $\mu_i^{KF}$ are simply $\mu_i^K$ shifted by a constant (see proposition 4.5) they follow the same distribution as $\mu_i^K$ except that its mean is shifted by the same constant. Similarly $\mu_i^{UA}$ and $\mu_i^{UF}$ are just $\mu_i^U$ shifted by a constant. So the results follow from proposition 4.2. □

For notational convenience, I introduce $q^{KA}$, $q^{KF}$, $q^{UA}$, and $q^{UF}$, defined by

$$\begin{aligned}
q^{KA} &= q^* + \frac{\varepsilon \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}, & q^{KF} &= q^* - \frac{\delta \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}, \\
q^{UA} &= q^* + \frac{\varepsilon \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}, & q^{UF} &= q^* - \frac{\delta \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}.
\end{aligned}$$

**Theorem 4.7.** *If $\varepsilon > 0$, $\delta > 0$, and $\sigma_{rv}^2 > 0$,*

$$\Pr\left(\mu_i^{KA} > q^*\right) < \Pr\left(\mu_i^{KF} > q^*\right) \text{ and } \Pr\left(\mu_i^{UA} > q^*\right) < \Pr\left(\mu_i^{UF} > q^*\right).$$

*Proof.* For the first inequality, note that

$$\Pr\left(\mu_i^{KA} > q^*\right) = 1 - \Phi\left(\frac{q^{KA} - \mu}{\sigma_K}\right) < 1 - \Phi\left(\frac{q^{KF} - \mu}{\sigma_K}\right) = \Pr\left(\mu_i^{KF} > q^*\right).$$

The equalities follow from the distributions of the posterior means established in proposition 4.6. The inequality follows from the fact that $\Phi$ is strictly increasing in its argument. By the same reasoning,

$$\Pr\left(\mu_i^{UA} > q^*\right) = 1 - \Phi\left(\frac{q^{UA} - \mu}{\sigma_U}\right) < 1 - \Phi\left(\frac{q^{UF} - \mu}{\sigma_U}\right) = \Pr\left(\mu_i^{UF} > q^*\right).$$

□

**Lemma D.4.**

$$\begin{aligned}
\Pr\left(A_i\right) &= p_{KA}\left(1 - \Phi\left(\frac{q^{KA} - \mu}{\sigma_K}\right)\right) + p_{KF}\left(1 - \Phi\left(\frac{q^{KF} - \mu}{\sigma_K}\right)\right) \\
&\quad + p_{UA}\left(1 - \Phi\left(\frac{q^{UA} - \mu}{\sigma_U}\right)\right) + p_{UF}\left(1 - \Phi\left(\frac{q^{UF} - \mu}{\sigma_U}\right)\right). \\
\mathbb{E}\left[q_i \mid A_i\right] &= \mu + \frac{\sigma_K}{\Pr\left(A_i\right)}\left(p_{KA}\phi\left(\frac{q^{KA} - \mu}{\sigma_K}\right) + p_{KF}\phi\left(\frac{q^{KF} - \mu}{\sigma_K}\right)\right) \\
&\quad + \frac{\sigma_U}{\Pr\left(A_i\right)}\left(p_{UA}\phi\left(\frac{q^{UA} - \mu}{\sigma_U}\right) + p_{UF}\phi\left(\frac{q^{UF} - \mu}{\sigma_U}\right)\right).
\end{aligned}$$

*Proof.* The expression for $\Pr(A_i)$ follows immediately from the distributions of the posterior means established in proposition 4.6.

To get an expression for $\mathbb{E}[q_i \mid A_i]$, consider first the average quality of scientist $i$'s paper given that it is accepted and given that scientist $i$ is in the group of scientists known to the editor that the editor is biased against. This average quality is

$$\mathbb{E}\left[q_i \mid \mu_i^{KA} > q^*\right] = \mathbb{E}\left[\mu_i^K \mid \mu_i^K > q^{KA}\right] = \mu + \sigma_K R\left(\frac{q^{KA} - \mu}{\sigma_K}\right),$$

where the first equality uses the fact that $\mu_i^{KA} > q^*$ is equivalent to $\mu_i^K > q^{KA}$ and then applies lemma D.1, and the second equality uses equation D.1. Similarly,

$$\begin{aligned}
\mathbb{E}\left[q_i \mid \mu_i^{KF} > q^*\right] &= \mu + \sigma_K R\left(\frac{q^{KF} - \mu}{\sigma_K}\right), \\
\mathbb{E}\left[q_i \mid \mu_i^{UA} > q^*\right] &= \mu + \sigma_U R\left(\frac{q^{UA} - \mu}{\sigma_U}\right), \\
\mathbb{E}\left[q_i \mid \mu_i^{UF} > q^*\right] &= \mu + \sigma_U R\left(\frac{q^{UF} - \mu}{\sigma_U}\right).
\end{aligned}$$

The average quality of accepted papers $\mathbb{E}[q_i \mid A_i]$ is a weighted sum of these expectations. The weights are given by the proportion of accepted papers that are written by a scientist in that particular group. For example, authors known to the editor that she is biased against form a $p_{KA} \Pr(\mu_i^{KA} > q^*)/\Pr(A_i)$ proportion of accepted papers. Hence

$$\begin{aligned}
\mathbb{E}\left[q_i \mid A_i\right] = &\frac{1}{\Pr\left(A_i\right)} p_{KA} \Pr\left(\mu_i^{KA} > q^*\right) \mathbb{E}\left[q_i \mid \mu_i^{KA} > q^*\right] \\
&+ \frac{1}{\Pr\left(A_i\right)} p_{KF} \Pr\left(\mu_i^{KF} > q^*\right) \mathbb{E}\left[q_i \mid \mu_i^{KF} > q^*\right] \\
&+ \frac{1}{\Pr\left(A_i\right)} p_{UA} \Pr\left(\mu_i^{UA} > q^*\right) \mathbb{E}\left[q_i \mid \mu_i^{UA} > q^*\right] \\
&+ \frac{1}{\Pr\left(A_i\right)} p_{UF} \Pr\left(\mu_i^{UF} > q^*\right) \mathbb{E}\left[q_i \mid \mu_i^{UF} > q^*\right] \\
= \mu &+ \frac{\sigma_K}{\Pr\left(A_i\right)} \left(p_{KA}\phi\left(\frac{q^{KA} - \mu}{\sigma_K}\right) + p_{KF}\phi\left(\frac{q^{KF} - \mu}{\sigma_K}\right)\right) \\
&+ \frac{\sigma_U}{\Pr\left(A_i\right)} \left(p_{UA}\phi\left(\frac{q^{UA} - \mu}{\sigma_U}\right) + p_{UF}\phi\left(\frac{q^{UF} - \mu}{\sigma_U}\right)\right). \qquad \square
\end{aligned}$$

# Bibliography

Melissa S. Anderson, Emily A. Ronning, Raymond De Vries, and Brian C. Martinson. Extending the Mertonian norms: Scientists' subscription to norms of research. *The Journal of Higher Education*, 81(3):366–393, 2010. ISSN 1538-4640. doi: 10.1353/jhe.0.0095. URL `https://muse.jhu.edu/journals/journal_of_higher_education/v081/81.3.anderson.html`.

Peter Arzberger, Peter Schroeder, Anne Beaulieu, Geof Bowker, Kathleen Casey, Leif Laaksonen, David Moorman, Paul Uhlir, and Paul Wouters. Promoting access to public research data for scientific, economic, and social development. *Data Science Journal*, 3:135–152, 2004. doi: 10.2481/dsj.3.135. URL `http://dx.doi.org/10.2481/dsj.3.135`.

Alain Aspect, Jean Dalibard, and Gérard Roger. Experimental test of Bell's inequalities using time-varying analyzers. *Physical Review Letters*, 49:1804–1807, Dec 1982. doi: 10.1103/PhysRevLett.49.1804. URL `http://link.aps.org/doi/10.1103/PhysRevLett.49.1804`.

Francis Bacon. *Novum Organon*. Longmans, London, 1620 [1858]. Translated by James Spedding et al. Accessed online on April 8, 2016. URL `http://en.wikisource.org/w/index.php?title=Novum_Organum/Book_I_(Spedding)&oldid=4460188`.

Francis Bacon. *New Atlantis*. Wikisource, 1626. Accessed online on April 8,

2016. URL https://en.wikisource.org/w/index.php?title=The_New_Atlantis&oldid=5709901.

Charles D. Bailey, Dana R. Hermanson, and Timothy J. Louwers. An examination of the peer review process in accounting journals. *Journal of Accounting Education*, 26(2):55–72, 2008a. ISSN 0748-5751. doi: 10.1016/j.jaccedu.2008.04.001. URL http://www.sciencedirect.com/science/article/pii/S0748575108000201.

Charles D. Bailey, Dana R. Hermanson, and James G. Tompkins. The peer review process in finance journals. *Journal of Financial Education*, 34: 1–27, 2008b. ISSN 0093-3961. URL http://www.jstor.org/stable/41948838.

Venkatesh Bala and Sanjeev Goyal. Learning from neighbours. *The Review of Economic Studies*, 65(3):595–621, 1998. doi: 10.1111/1467-937X.00059. URL http://restud.oxfordjournals.org/content/65/3/595.full.pdf+html.

Abhijit V. Banerjee. A simple model of herd behavior. *The Quarterly Journal of Economics*, 107(3):797–817, 1992. ISSN 00335533. URL http://www.jstor.org/stable/2118364.

Siddhartha Banerjee, Ashish Goel, and Anilesh Kollagunta Krishnaswamy. Re-incentivizing discovery: Mechanisms for partial-progress sharing in research. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, EC '14, pages 149–166, New York, 2014. ACM. ISBN 978-1-4503-2565-3. doi: 10.1145/2600057.2602888. URL http://dx.doi.org/10.1145/2600057.2602888.

C. Glenn Begley and Lee M. Ellis. Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533, Mar 2012. doi: 10.1038/483531a. URL http://dx.doi.org/10.1038/483531a.

Damien Besancenot, Kim V. Huynh, and Joao R. Faria. Search and research: the influence of editorial boards on journals' quality. *Theory and Decision*, 73(4):687–702, 2012. ISSN 0040-5833. doi: 10.1007/s11238-012-9314-7. URL http://dx.doi.org/10.1007/s11238-012-9314-7.

Cristina Bicchieri. Methodological rules as conventions. *Philosophy of the Social Sciences*, 18(4):477–495, 1988. ISSN 0048-3931. doi: 10.1177/004839318801800403. URL http://pos.sagepub.com/content/18/4/477.short.

Cristina Bicchieri. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press, Cambridge, 2006. ISBN 9780521574907.

Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5):992–1026, 1992. ISSN 00223808. URL http://www.jstor.org/stable/2138632.

Rebecca M. Blank. The effects of double-blind versus single-blind reviewing: Experimental evidence from The American Economic Review. *The American Economic Review*, 81(5):1041–1067, 1991. ISSN 00028282. URL http://www.jstor.org/stable/2006906.

David Bloor. *Knowledge and Social Imagery*. University Of Chicago Press, Chicago, second edition, 1991. ISBN 0226060977.

Christine L. Borgman. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63(6):1059–1078, 2012. ISSN 1532-2890. doi: 10.1002/asi.22634. URL http://dx.doi.org/10.1002/asi.22634.

Richard N. Boyd. On the current status of the issue of scientific realism.

*Erkenntnis*, 19(1):45–90, 1983. ISSN 01650106. URL http://www.jstor.org/stable/20010835.

Thomas Boyer. Is a bird in the hand worth two in the bush? Or, whether scientists should publish intermediate results. *Synthese*, 191(1): 17–35, 2014. ISSN 0039-7857. doi: 10.1007/s11229-012-0242-4. URL http://dx.doi.org/10.1007/s11229-012-0242-4.

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

Liam Kofi Bright. Against candidate quality. Manuscript, 2015. URL https://www.academia.edu/11673059/Against_Candidate_Quality.

Liam Kofi Bright. On fraud. *Philosophical Studies*, forthcoming. ISSN 1573-0883. doi: 10.1007/s11098-016-0682-7. URL http://dx.doi.org/10.1007/s11098-016-0682-7.

Liam Kofi Bright, Daniel Malinsky, and Morgan Thompson. Causally interpreting intersectionality theory. *Philosophy of Science*, 83(1):60–81, 2016. ISSN 00318248. URL http://www.jstor.org/stable/10.1086/684173.

Justin Bruner and Cailin O'Connor. Power, bargaining, and collaboration. In Thomas Boyer-Kassem, Conor Mayo-Wilson, and Michael Weisberg, editors, *Scientific Collaboration and Collective Knowledge*. Oxford University Press, Oxford, forthcoming. URL http://philpapers.org/rec/BRUPBA-2.

Justin P. Bruner. Policing epistemic communities. *Episteme*, 10:403–416, Dec 2013. ISSN 1750-0117. doi: 10.1017/epi.2013.34. URL http://journals.cambridge.org/article_S1742360013000348.

John M. Budd, MaryEllen Sievert, and Tom R. Schultz. Phenomena of retraction: Reasons for retraction and citations to the publications. *Journal of the American Medical Association*, 280(3):296–297, 1998. doi: 10.1001/jama.280.3.296. URL `http://dx.doi.org/10.1001/jama.280.3.296`.

Amber E. Budden, Tom Tregenza, Lonnie W. Aarssen, Julia Koricheva, Roosa Leimu, and Christopher J. Lortie. Double-blind review favours increased representation of female authors. *Trends in Ecology & Evolution*, 23(1):4–6, 2008. ISSN 0169-5347. doi: 10.1016/j.tree.2007.07.008. URL `http://www.sciencedirect.com/science/article/pii/S0169534707002704`.

Rudolf Carnap. Empiricism, semantics, and ontology. *Revue Internationale de Philosophie*, 4:20–40, 1950. ISSN 00488143.

Gerald Carter. Goals of science vs goals of scientists (& a love letter to PLOS One). Accessed online on April 18, 2016. URL `http://socialbat.org/2015/08/12/goals-of-science-vs-goals-of-scientists-a-love-letter-for-plos-one/`, Aug 2015.

Stephen J. Ceci and Wendy M. Williams. Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences*, 108(8):3157–3162, 2011. doi: 10.1073/pnas.1014871108. URL `http://www.pnas.org/content/108/8/3157.abstract`.

Kenneth E. Clark. *America's Psychologists: A Survey of a Growing Profession*. American Psychological Association, Washington, 1957.

Frank Close. *Too Hot to Handle: The Race for Cold Fusion*. Princeton University Press, Princeton, 1991.

Lorraine Code. *What Can She Know? Feminist Theory and the Construction of Knowledge*. Cornell University Press, Ithaca, 1991. ISBN 0801497205.

Jonathan R. Cole and Stephen Cole. Measuring the quality of sociological research: Problems in the use of the "Science Citation Index". *The American Sociologist*, 6(1):23–29, 1971. ISSN 00031232. URL http://www.jstor.org/stable/27701705.

Jonathan R. Cole and Stephen Cole. *Social Stratification in Science*. University of Chicago Press, Chicago, 1973. ISBN 0226113388.

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

Stephen Cole and Jonathan R. Cole. Visibility and the structural bases of awareness of scientific research. *American Sociological Review*, 33(3): 397–413, 1968. ISSN 00031224. URL http://www.jstor.org/stable/2091914.

H. M. Collins. The place of the 'core-set' in modern science: Social contingency with methodological propriety in science. *History of Science*, 19(1):6–19, 1981. doi: 10.1177/007327538101900102. URL http://hos.sagepub.com/content/19/1/6.short.

Patricia Hill Collins and Valerie Chepp. Intersectionality. In Georgina Waylen, Karen Celis, Johanna Kantola, and S. Laurel Weldon, editors, *The Oxford Handbook of Gender and Politics*, chapter 2, pages 57–87. Oxford University Press, Oxford, 2013. ISBN 0199751455.

Rick Crandall. Editorial responsibilities in manuscript review. *Behavioral and Brain Sciences*, 5:207–208, Jun 1982. ISSN 1469-1825. doi: 10.1017/S0140525X00011316. URL http://journals.cambridge.org/article_S0140525X00011316.

Diana Crane. The gatekeepers of science: Some factors affecting the selection of articles for scientific journals. *The American Sociologist*, 2(4):195–201, 1967. ISSN 00031232. URL `http://www.jstor.org/stable/27701277`.

Partha Dasgupta and Paul A. David. Toward a new economics of science. *Research Policy*, 23(5):487–521, 1994. ISSN 0048-7333. doi: 10.1016/0048-7333(94)01002-1. URL `http://www.sciencedirect.com/science/article/pii/0048733394010021`.

Morris H. DeGroot. *Optimal Statistical Decisions*. John Wiley & Sons, New Jersey, 2004.

Pierre Duhem. *La Théorie Physique: son Objet et sa Structure*. Jules Vuillemin, Paris, 1906.

Kenny Easwaran. Probabilistic proofs and transferability. *Philosophia Mathematica*, 17(3):341–362, 2009. doi: 10.1093/philmat/nkn032. URL `http://philmat.oxfordjournals.org/content/17/3/341.abstract`.

Glenn Ellison. Evolving standards for academic publishing: A q-r theory. *Journal of Political Economy*, 110(5):994–1034, 2002a. ISSN 00223808. URL `http://www.jstor.org/stable/10.1086/341871`.

Glenn Ellison. The slowdown of the economics publishing process. *Journal of Political Economy*, 110(5):947–993, 2002b. ISSN 00223808. URL `http://www.jstor.org/stable/10.1086/341868`.

Daniele Fanelli. Do pressures to publish increase scientists' bias? An empirical support from US states data. *PLoS ONE*, 5(4):e10271, Apr 2010. doi: 10.1371/journal.pone.0010271. URL `http://dx.doi.org/10.1371/journal.pone.0010271`.

João Ricardo Faria. The game academics play: Editors versus authors. *Bulletin of Economic Research*, 57(1):1–12, 2005. ISSN 1467-8586. doi:

10.1111/j.1467-8586.2005.00212.x. URL http://dx.doi.org/10.1111/j.1467-8586.2005.00212.x.

Paul Feyerabend. *Against Method*. New Left Books, London, 1975.

Miranda Fricker. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press, Oxford, 2007.

Michael Friedman. Truth and confirmation. *The Journal of Philosophy*, 76(7):361–382, 1979. ISSN 0022362X. URL http://www.jstor.org/stable/2025452.

Steve Fuller. *Social Epistemology*. Indiana University Press, Bloomington, second edition, 2002.

Douglas Gale and Shachar Kariv. Bayesian learning in social networks. *Games and Economic Behavior*, 45(2):329–346, 2003. ISSN 0899-8256. doi: 10.1016/S0899-8256(03)00144-1. URL http://www.sciencedirect.com/science/article/pii/S0899825603001441.

Robert D. Gordon. Values of Mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *The Annals of Mathematical Statistics*, 12(3):364–366, 1941. ISSN 00034851. URL http://www.jstor.org/stable/2235868.

Simon J. Goring, Kathleen C. Weathers, Walter K. Dodds, Patricia A. Soranno, Lynn C. Sweet, Kendra S. Cheruvelil, John S. Kominoski, Janine Rüegg, Alexandra M. Thorn, and Ryan M. Utz. Improving the culture of interdisciplinary collaboration in ecology by expanding measures of success. *Frontiers in Ecology and the Environment*, 12(1):39–47, Feb 2014. ISSN 1540-9295. doi: 10.1890/120370. URL http://dx.doi.org/10.1890/120370.

Remco Heesen. How much evidence should one collect? *Philosophical Studies*, 172(9):2299–2313, 2015. ISSN 0031-8116. doi: 10.1007/s11098-014-0411-z. URL http://dx.doi.org/10.1007/s11098-014-0411-z.

Remco Heesen. Academic superstars: Competent or lucky? *Synthese*, forthcoming. doi: 10.1007/s11229-016-1146-5. URL http://dx.doi.org/10.1007/s11229-016-1146-5.

Martin Heintzelman and Diego Nocetti. Where should we submit our manuscript? An analysis of journal submission strategies. *The B.E. Journal of Economic Analysis & Policy*, 9(1), Sep 2009. ISSN 1935-1682. doi: 10.2202/1935-1682.2340. URL http://dx.doi.org/10.2202/1935-1682.2340.

Carl G. Hempel. The theoretician's dilemma: A study in the logic of theory construction. In Herbert Feigl, Michael Scriven, and Grover Maxwell, editors, *Minnesota Studies in the Philosophy of Science*, volume 2, pages 37–98. University of Minnesota Press, Minneapolis, 1958.

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

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

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

John C. Huber and Roland Wagner-Döbler. Scientific production: A statistical analysis of authors in mathematical logic. *Scientometrics*, 50(2): 323–337, 2001a. ISSN 0138-9130. doi: 10.1023/A:1010581925357. URL http://dx.doi.org/10.1023/A%3A1010581925357.

John C. Huber and Roland Wagner-Döbler. Scientific production: A statistical analysis of authors in physics, 1800-1900. *Scientometrics*, 50(3): 437–453, 2001b. ISSN 0138-9130. doi: 10.1023/A:1010558714879. URL http://dx.doi.org/10.1023/A%3A1010558714879.

John R. Huizenga. *Cold Fusion: The Scientific Fiasco of the Century*. Oxford University Press, Oxford, second edition, 1993.

David L. Hull. *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science*. University of Chicago Press, Chicago, 1988. ISBN 0226360504.

Simon M. Huttegger, Brian Skyrms, and Kevin J. S. Zollman. Probe and adjust in information transfer games. *Erkenntnis*, 79(4):835–853, 2014. ISSN 0165-0106. doi: 10.1007/s10670-013-9467-y. URL http://dx.doi.org/10.1007/s10670-013-9467-y.

John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, Aug 2005. doi: 10.1371/journal.pmed.0020124. URL http://dx.doi.org/10.1371/journal.pmed.0020124.

Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*, volume 1. John Wiley & Sons, New York, second edition, 1994.

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

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

Philip Kitcher. *The Advancement of Science: Science without Legend, Objectivity without Illusions*. Oxford University Press, Oxford, 1993. ISBN 0195046285.

Thomas S. Kuhn. *The Structure of Scientific Revolutions*. The University of Chicago Press, Chicago, 1962.

David N. Laband. Publishing favoritism: A critique of department rankings based on quantitative publishing performance. *Southern Economic Journal*, 52(2):510–515, 1985. ISSN 00384038. URL `http://www.jstor.org/stable/1059636`.

David N. Laband and Michael J. Piette. Favoritism versus search for good papers: Empirical evidence regarding the behavior of journal editors. *Journal of Political Economy*, 102(1):194–203, 1994a. ISSN 00223808. URL `http://www.jstor.org/stable/2138799`.

David N. Laband and Michael J. Piette. Does the "blindness" of peer review influence manuscript selection efficiency? *Southern Economic Journal*, 60(4):896–906, 1994b. ISSN 00384038. URL `http://www.jstor.org/stable/1060428`.

Imre Lakatos. Criticism and the methodology of scientific research programmes. *Proceedings of the Aristotelian Society*, 69:149–186, 1968. ISSN 00667374. URL `http://www.jstor.org/stable/4544774`.

Bruno Latour and Steve Woolgar. *Laboratory Life: The Construction of Scientific Facts*. Princeton University Press, Princeton, second edition, 1986.

Carole J. Lee. Revisiting current causes of women's underrepresentation in science. In Jennifer Saul and Michael Brownstein, editors, *Implicit Bias and Philosophy Volume 1: Metaphysics and Epistemology*. Oxford University Press, Oxford, forthcoming.

Carole J. Lee, Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1):2–17, 2013. ISSN 1532-2890. doi: 10.1002/asi.22784. URL http://dx.doi.org/10.1002/asi.22784.

Hannes Leitgeb. Scientific philosophy, mathematical philosophy, and all that. *Metaphilosophy*, 44(3):267–275, 2013. ISSN 1467-9973. doi: 10.1111/meta.12029. URL http://dx.doi.org/10.1111/meta.12029.

D. Lindsey. Using citation counts as a measure of quality in science: Measuring what's measurable rather than what's valid. *Scientometrics*, 15 (3–4):189–203, 1989. ISSN 0138-9130. doi: 10.1007/BF02017198. URL http://dx.doi.org/10.1007/BF02017198.

Karen Seashore Louis, Lisa M. Jones, and Eric G. Campbell. Macroscope: Sharing in science. *American Scientist*, 90(4):304–307, 2002. ISSN 00030996. URL http://www.jstor.org/stable/27857685.

Bruce Macfarlane and Ming Cheng. Communism, universalism and disinterestedness: Re-examining contemporary support among academics for Merton's scientific norms. *Journal of Academic Ethics*, 6(1):67–78, 2008. ISSN 1570-1727. doi: 10.1007/s10805-008-9055-y. URL http://dx.doi.org/10.1007/s10805-008-9055-y.

Christopher D. Mackie. *Canonizing Economic Theory: How Theories and Ideas Are Selected in Economics*. M. E. Sharpe, New York, 1998. ISBN 9780765602848.

Conor Mayo-Wilson, Kevin J. S. Zollman, and David Danks. The independence thesis: When individual and social epistemology diverge. *Philosophy of Science*, 78(4):653–677, 2011. ISSN 00318248. URL `http://www.jstor.org/stable/10.1086/661777`.

Conor Mayo-Wilson, Kevin J. S. Zollman, and David Danks. Wisdom of crowds versus groupthink: learning in groups and in isolation. *International Journal of Game Theory*, 42(3):695–723, 2013. ISSN 0020-7276. doi: 10.1007/s00182-012-0329-7. URL `http://dx.doi.org/10.1007/s00182-012-0329-7`.

Marshall H. Medoff. Editorial favoritism in economics? *Southern Economic Journal*, 70(2):425–434, 2003. ISSN 00384038. URL `http://www.jstor.org/stable/3648979`.

Robert K. Merton. A note on science and democracy. *Journal of Legal and Political Sociology*, 1(1–2):115–126, 1942. Reprinted in Merton (1973, chapter 13).

Robert K. Merton. Priorities in scientific discovery: A chapter in the sociology of science. *American Sociological Review*, 22(6):635–659, 1957. ISSN 00031224. URL `http://www.jstor.org/stable/2089193`. Reprinted in Merton (1973, chapter 14).

Robert K. Merton. Singletons and multiples in scientific discovery: A chapter in the sociology of science. *Proceedings of the American Philosophical Society*, 105(5):470–486, 1961. ISSN 0003049X. URL `http://www.jstor.org/stable/985546`. Reprinted in Merton (1973, chapter 16).

Robert K. Merton. Behavior patterns of scientists. *The American Scholar*, 38 (2):197–225, 1969. ISSN 00030937. URL `http://www.jstor.org/stable/41209646`. Reprinted in Merton (1973, chapter 15).

Robert K. Merton. *The Sociology of Science: Theoretical and Empirical Investigations*. The University of Chicago Press, Chicago, 1973. ISBN 0226520919.

Charles W. Mills. White ignorance. In Shannon Sullivan and Nancy Tuana, editors, *Race and Epistemologies of Ignorance*, chapter 1, pages 13–38. State University of New York Press, 2007. ISBN 0791471012.

Corinne A. Moss-Racusin, John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, 109(41): 16474–16479, 2012. doi: 10.1073/pnas.1211286109. URL `http://www.pnas.org/content/109/41/16474.abstract`.

Leif D. Nelson, Joseph P. Simmons, and Uri Simonsohn. Let's publish fewer papers. *Psychological Inquiry*, 23(3):291–293, 2012. doi: 10.1080/1047840X.2012.705245. URL `http://dx.doi.org/10.1080/1047840X.2012.705245`.

James R. Norris. *Markov Chains*. Cambridge University Press, Cambridge, 1998. URL `http://dx.doi.org/10.1017/CBO9780511810633`.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251), Aug 2015. doi: 10.1126/science.aac4716. URL `http://www.sciencemag.org/content/349/6251/aac4716.abstract`.

Harold Pashler and Eric-Jan Wagenmakers. Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7(6):528–530, 2012. doi: 10.1177/1745691612465253. URL `http://pps.sagepub.com/cgi/content/short/7/6/528`.

Michael J. Piette and Kevin L. Ross. A study of the publication of scholarly

output in economics journals. *Eastern Economic Journal*, 18(4):429–436, 1992. ISSN 00945056. URL http://www.jstor.org/stable/40325474.

Heather Piwowar. Altmetrics: Value all research products. *Nature*, 493 (7431):159, Jan 2013. ISSN 1476-4687. doi: 10.1038/493159a. URL http://dx.doi.org/10.1038/493159a.

Robert Pool. Fusion followup: Confusion abounds. *Science*, 244(4900):27–29, 1989. doi: 10.1126/science.244.4900.27. URL http://www.sciencemag.org/content/244/4900/27.short.

Karl Popper. *The Logic of Scientific Discovery*. Hutchinson, London, 1959.

Derek J. de Solla Price. *Little Science, Big Science*. Columbia University Press, New York, 1963.

Derek J. de Solla Price. Networks of scientific papers. *Science*, 149(3683): 510–515, 1965. ISSN 00368075. URL http://www.jstor.org/stable/1716232.

Florian Prinz, Thomas Schlange, and Khusru Asadullah. Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, 10(9):712, Sep 2011. doi: 10.1038/nrd3439-c1. URL http://dx.doi.org/10.1038/nrd3439-c1.

W. V. Quine. Two dogmas of empiricism. *The Philosophical Review*, 60 (1):20–43, 1951. ISSN 00318108. URL http://www.jstor.org/stable/2181906.

David B. Resnik. Openness versus secrecy in scientific research. *Episteme*, 2: 135–147, Oct 2006. ISSN 1750-0117. doi: 10.3366/epi.2005.2.3.135. URL http://journals.cambridge.org/article_S174236000000037X.

Robert Rosenthal. The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, 86(3):638–641, 1979. doi: 10.1037/0033-2909.86.

3.638. URL http://psycnet.apa.org/doi/10.1037/0033-2909.86.3.638.

Jennifer Saul. Implicit bias, stereotype threat, and women in philosophy. In Katrina Hutchison and Fiona Jenkins, editors, *Women in Philosophy: What Needs to Change?*, chapter 2, pages 39–60. Oxford University Press, Oxford, 2013.

Daniel L. Sherrell, Joseph F. Hair, Jr., and Mitch Griffin. Marketing academicians' perceptions of ethical research and publishing behavior. *Journal of the Academy of Marketing Science*, 17(4):315–324, 1989. ISSN 0092-0703. doi: 10.1007/BF02726642. URL http://dx.doi.org/10.1007/BF02726642.

Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366, 2011. doi: 10.1177/0956797611417632. URL http://pss.sagepub.com/content/22/11/1359.abstract.

Kenneth J. Smith and Robert F. Dombrowski. An examination of the relationship between author-editor connections and subsequent citations of auditing research articles. *Journal of Accounting Education*, 16 (3–4):497–506, 1998. ISSN 0748-5751. doi: 10.1016/S0748-5751(98)00019-0. URL http://www.sciencedirect.com/science/article/pii/S0748575198000190.

Richard Smith. Peer review: a flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine*, 99(4):178–182, 2006. URL http://jrs.sagepub.com/content/99/4/178.short.

Patricia A. Soranno, Kendra S. Cheruvelil, Kevin C. Elliott, and Georgina M. Montgomery. It's good to share: Why environmental scientists' ethics

are out of date. *BioScience*, 65(1):69–73, 2015. doi: 10.1093/biosci/biu169. URL http://bioscience.oxfordjournals.org/content/65/1/69.abstract.

Jan Sprenger and Remco Heesen. The bounded strength of weak expectations. *Mind*, 120(479):819–832, 2011. ISSN 00264423. URL http://www.jstor.org/stable/41494380.

Rhea E. Steinpreis, Katie A. Anders, and Dawn Ritzke. The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A national empirical study. *Sex Roles*, 41(7–8):509–528, 1999. ISSN 0360-0025. doi: 10.1023/A:1018839203698. URL http://dx.doi.org/10.1023/A:1018839203698.

Brandon D. Stewart and B. Keith Payne. Bringing automatic stereotyping under control: Implementation intentions as efficient means of thought control. *Personality and Social Psychology Bulletin*, 34(10):1332–1345, 2008. doi: 10.1177/0146167208321269. URL http://psp.sagepub.com/content/34/10/1332.abstract.

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

Michael Strevens. Herding and the quest for credit. *Journal of Economic Methodology*, 20(1):19–34, 2013. doi: 10.1080/1350178X.2013.774849. URL http://dx.doi.org/10.1080/1350178X.2013.774849.

Michael Strevens. Scientific sharing: Communism and the social contract. In Thomas Boyer-Kassem, Conor Mayo-Wilson, and Michael Weisberg, editors, *Scientific Collaboration and Collective Knowledge*. Oxford University Press, Oxford, forthcoming. URL http://www.strevens.org/research/scistruc/communicans.shtml.

Athina Tatsioni, Nikolaos G. Bonitsis, and John P. A. Ioannidis. Persistence of contradicted claims in the literature. *Journal of the American Medical Association*, 298(21):2517–2526, 2007. doi: 10.1001/jama.298.21.2517. URL http://dx.doi.org/10.1001/jama.298.21.2517.

Carol Tenopir, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. Data sharing by scientists: Practices and perceptions. *PLoS ONE*, 6(6):e21101, Jun 2011. doi: 10.1371/journal.pone.0021101. URL http://dx.doi.org/10.1371/journal.pone.0021101.

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

Eric Luis Uhlmann and Geoffrey L. Cohen. "I think it, therefore it's true": Effects of self-perceived objectivity on hiring discrimination. *Organizational Behavior and Human Decision Processes*, 104(2):207–223, 2007. ISSN 0749-5978. doi: 10.1016/j.obhdp.2007.07.001. URL http://www.sciencedirect.com/science/article/pii/S0749597807000611.

Virginia Valian. *Why So Slow? The Advancement of Women*. MIT Press, Cambridge, 1999. ISBN 9780262720311.

Bas C. van Fraassen. *The Scientific Image*. Oxford University Press, Oxford, 1980.

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

Christine Wennerås and Agnes Wold. Nepotism and sexism in peer-review. *Nature*, 387(6631):341–343, May 1997. ISSN 0028-0836. doi: 10.1038/387341a0. URL http://dx.doi.org/10.1038/387341a0.

Alan J. Ziobrowski and Karen M. Gibler. Factors academic real estate authors consider when choosing where to submit a manuscript for publication. *Journal of Real Estate Practice and Education*, 3(1):43–54, 2000. ISSN 1521-4842. URL http://ares.metapress.com/content/1762151051KM2227.

Kevin J. S. Zollman. Optimal publishing strategies. *Episteme*, 6:185–199, Jun 2009. ISSN 1750-0117. doi: 10.3366/E174236000900063X. URL http://journals.cambridge.org/article_S1742360000001283.

Kevin J. S. Zollman. The epistemic benefit of transient diversity. *Erkenntnis*, 72(1):17–35, 2010. ISSN 01650106. URL http://www.jstor.org/stable/20642278.

Kevin J. S. Zollman. Modeling the social consequences of testimonial norms. *Philosophical Studies*, 172(9):2371–2383, 2015. ISSN 1573-0883. doi: 10.1007/s11098-014-0416-7. URL http://dx.doi.org/10.1007/s11098-014-0416-7.

Harriet Zuckerman and Robert K. Merton. Patterns of evaluation in science: Institutionalisation, structure and functions of the referee system. *Minerva*, 9(1):66–100, 1971. ISSN 0026-4695. doi: 10.1007/BF01553188. URL http://dx.doi.org/10.1007/BF01553188. Reprinted in Merton (1973, chapter 21).